UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

FACULTAD DE ECONOMÍA

"PRONÓSTICO DE LA TASA DE DESEMPLEO PARA PERSONAS CON EDUCACIÓN MEDIA SUPERIOR Y SUPERIOR EN MÉXICO PARA 2018 UTILIZANDO SIMULACIÓN"

TESIS

PARA OBTENER EL TITULO DE:

LICENCIADO EN ACTUARIA

PRESENTA:

JULIO TONATIUH BALCAZAR LOPEZ

ASESOR:

M. EN E.S. R. y M. EMILIO DAVID OLVERA REBOLLEDO

REVISORES:

Dra. EN Hum. MARGARITA JOSEFINA HOLGUIN GARCIA

M. EN Eco. LEOBARDO DE JESUS ALMONTE

TOLUCA, ESTADO DE MEXICO **MARZO 2020**

INDICE

Introducción

Una de las preocupaciones de la economía que más tienen eco en la sociedad es el desempleo, ya que afecta directamente a todo el ambiente político y social debido a que uno de los objetivos principales de un país es proveer un nivel satisfactorio de empleo para la población. Este nivel de empleo es satisfactorio cuando todos o casi todos sus miembros tienen un empleo. El desempleo se produce cuando ocurre un desequilibrio entre la oferta de mano de obra y la demanda de trabajo. Si hay más mano de obra disponible para trabajar y pocas oportunidades de empleo debido a la escasa creación de éstos, se crea un excedente de recursos humanos que van a ingresar a las filas de desempleados. (Romero, 2008)

Los empleos vinculan a las personas con la sociedad y la economía en las que viven. El acceso a un trabajo seguro, productivo y remunerado de manera justa es un factor fundamental para la autoestima de las personas y las familias, que les afirma su sentimiento de pertenencia a una comunidad, y les permite hacer una contribución productiva. El cambio hacia un desarrollo incluyente y sostenible no será posible si se niega a millones de personas la oportunidad de ganarse la vida en condiciones dignas y equitativas.

De acuerdo con Ruiz & Ordaz (2011) una de las causas más directas de la pobreza en las sociedades modernas es la falta de oportunidades para la población económicamente activa de encontrar un empleo con la suficiente remuneración en una economía que tiene en la tecnificación una de sus principales características. El crecimiento económico no siempre es un indicador confiable para medir la utilización plena de la mano de obra disponible que es usualmente abundante y poco calificada en los países menos desarrollados como México.

Es importante explorar los datos recientes en México para con ello poder determinar ciertos factores temporales que afectan el desempleo en personas semi y profesionalizadas.

El problema del desempleo en México es la unión de dos grandes tópicos, la elevada cantidad de personas en edad laboral además de la poca capacidad del sistema económico para generar los suficientes empleos adecuados para la demanda.

Como bien dicen Ruiz Nápoles & Ordaz Díaz (2011) se suponía que siendo México un país relativamente abundante en mano de obra, una vez abierta su economía, ésta se especializaría con ventaja en la producción de bienes intensivos en mano de obra, por las diferencias salariales y de dotación relativa con Estados Unidos. Dichos bienes al ser manufacturados, tendrían un efecto de arrastre importante en la generación de empleos en las ramas proveedoras locales. Estas ventajas se manifestarían en una mayor exportación de bienes intensivos en mano de obra. Un factor adicional generador de empleos estimulado por las reformas fue la instalación de empresas maquiladoras, ya no restringida a la zona norte del país como inicialmente en los sesenta y setenta.

La expansión de la informalidad o subempleo es otro de los problemas que México comparte con la región Latinoamericana e incluso con otros países desarrollados y en desarrollo.

En el informe de la Organización Internacional del Trabajo (2016) se especifica como la economía ha crecido en los últimos años, más de lo previsto, sin embargo, si se mantienen las políticas actuales habrán problemas importantes para las empresas y trabajadores.

Sin embargo, aunque pareciera que se crean condiciones para que las personas con Educación Media Superior y Superior tengan mejores oportunidades, esto no sucede así. Burgos Flores y López Montes (2010) especifican, lo característico en los últimos años fue la aceleración del crecimiento en el ritmo de la educación media superior y superior.

Además del fenómeno del desempleo, han aparecido una serie de distorsiones en el mercado laboral de profesionistas, tales como: la ocupación de puestos que no requieren de estudios (sobreeducación); la baja coincidencia de los conocimientos y habilidades adquiridos en las instituciones de educación superior y las funciones desempeñadas (desfase de conocimientos) y los correspondientes bajos niveles salariales, entre otros.

Para el caso de México, varios estudios arrojan resultados que describen la intensidad de esta problemática. Un primer estudio en la década de 1990-2000 encontró que, cerca de las dos terceras partes de los profesionistas mexicanos se emplean en puestos acordes a la profesión, los cuales es probable que apliquen los conocimientos y habilidades adquiridas en las instituciones de educación superior y que el otro grupo podría estar ocupando puestos que no requieren de educación superior.

Lo anterior configura una situación preocupante del mercado laboral para las personas con educación media superior y superior en el país, además de la adecuada inserción en el mercado laboral por parte de los egresados es una de sus dimensiones más importantes.

El presente trabajo se plantea como objetivo analizar los datos mensuales del indicador de desempleo brindado por el banco de información económica de INEGI y con ello tener estimadores de la tasa de desempleo calculados con un alto grado de confianza.

Al finalizar el trabajo se contará con los valores proyectados de todos los meses de 2018 lo cual es una base fundamental para que se pueda trabajar, teniéndolos se pueden contrastar las reformas que en su momento aplique el gobierno además de analizar si los proyectos de inversión disminuyen de manera considerable las tasas de desempleo estudiadas.

La tasa de desempleo refleja más el bienestar de las familias que la actividad económica. Es de las variables que más puede influenciar el comportamiento de los mercados si es que se aparta mucho de la expectativa.

En cambio, en México la mayoría de los analistas le otorga muy poca importancia. Prácticamente no hay reacción de los mercados, aún en el caso de que la cifra reportada resulta ser muy diferente a la esperada. Únicamente los medios le dan un lugar en las primeras planas, pero siempre y cuando la tasa sea mayor al mes (o año) anterior.

Es por ello que con los resultados esperados se tendrá una base fundamental para que se pueda trabajar, ya sea dentro del gobierno o para las finanzas.

De acuerdo a la tendencia general, los datos del desempleo en educación media superior y superior en México durante los últimos años tienen una gran sensibilidad a los cambios

políticos, es decir, la dependencia de los datos es temporal pero sólo para los años más recientes por lo cual el pronóstico será muy inestable.

Sin embargo, teniendo un proceso donde se especifica la metodología utilizada para modelar las series de tiempo además de detallar el contexto del desempleo en México especificando la metodología de medición del país, dará las bases necesarias para generar el modelo ARIMA modificado más confiable basado en la simulación para pronosticar a un alto grado de confianza.

Para analizar el desempleo de manera correcta se tiene que utilizar una metodología de investigación científica económica ya que se trata de encontrar la solución científica con bases matemáticas, en este caso, para pronosticar una serie histórica. Además, el resultado del presente trabajo brindará información acerca de una variable ampliamente estudiada en la economía.

Lo primero que se realizará es un análisis cualitativo del desempleo en personas con educación media superior y superior durante los últimos años en México, una vez realizado esto, se podrá determinar una tendencia histórica del estimador además de limitar el problema en estudio.

Ya que se tienen las nociones básicas además del panorama definido por los principales teóricos, es necesario adentrarse en la serie histórica específica del desempleo en educación media superior y superior. Se definirán los conceptos claves para entender la tendencia de los datos y con ello tener la base para desarrollar el problema específico con la información obtenida de fuentes confiables.

Por último, se tomarán los resultados y se les aplicarán las técnicas necesarias y marcadas en la teoría series de tiempo y de simulación para poder determinar un pronóstico con alto grado de confianza. Una vez concluido este proceso se remarcará la importancia de los resultados obtenidos además de las interpretaciones teóricas de los mismos.

I. Modelos de regresión y contexto del desempleo en herramientas matemáticas

La forma que se ha adoptado para medir el desempleo ha sido un porcentaje en base a la población activa del país. Generalmente, el porcentaje de desempleo se toma tiempo en tener información suficiente para pronostico debido a sus fuentes históricas amplias. A causa de ser de los fenómenos que se estudian más de una economía, se tienen registros amplios acerca de su evaluación en el tiempo. Esto es de gran utilidad porque se puede analizar sí el tiempo en sí es un factor para el desarrollo de esta variable económica, además de tener influencia de otros elementos relacionados con la periodicidad.

Una forma base para entender la series de tiempo es la regresión, pero esta es encontrar una función con variables independientes para predecir una variable dependiente que podría ser el precio de una acción, el valor del PIB, etc. Una forma de explicar la regresión sería por ejemplo, utilizar el precio del dólar-peso para evaluar la disminución o aumento de las exportaciones. Pero para poder predecir de manera correcta es necesario tener una serie de datos históricos de todos los días de registro del cierre del precio del dólar.

Una forma de explicar de una mejor manera podría ser incluir más variables independientes al modelo como el gasto público puede influir de cierta manera en el incentivo que pueden tener los empresarios a invertir y exportar productos.

En general, se puede decir que una variable puede ser afectada por variables independientes, pero esto no siempre define que el modelo sea el mejor, en cambio, se puede decir que esta variable solamente depende del tiempo en sí para saber cuál será su comportamiento. Es decir, el valor de las exportaciones no depende de ninguna variable externa sino solamente del valor de estas en los días anteriores.

El caso de las variables en las que solo depende de los valores anteriores a esta se utiliza un método llamado autoregresión AR (posteriormente se explicará la relación con el modelo ARIMA).

1.1 Series modelables

Los métodos que se explicarán más adelante están aplicados a series con las siguientes características:

- Media constante: Es decir el valor medio de la serie no debe cambiar de manera drástica en el tiempo.
- Varianza constante u homocedasticidad: La anchura de los datos de la serie no deben de variar significativamente en tramos marcados.

Una serie de tiempo tiene generalmente variaciones en la media además de varianza no constante en los datos y ciclos, esto hay que arreglarlo para poder empezar a trabajar, más tarde se ahondará en este tema.

1.2 Proceso autoregresivo AR

El proceso autorregresivo tiene motivación directa, es simplemente una ecuación que expresa diferencias estocásticas, un modelo matemático sencillo en el que el valor actual de la serie está linealmente relacionado con sus valores en el pasado, más un choque estocástico aditivo. Box y Jenkins (1970) definen la metodología a utilizar para analizar las series de tiempo. Mostrado a continuación.

El proceso autoregresivo general de orden p, AR(p) tiene la siguiente forma:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t \qquad (1.1)$$

Donde el proceso (ε_t) es un ruido blanco con varianza σ^2 y media cero. Los ε_t se llaman innovaciones, ya que representan la nueva información que aparece en cada instante.

El valor actual de la serie es una función lineal de los valores anteriores y la innovación actual. El proceso se puede expresar como una función lineal de todas las innovaciones.

Si se emplea el operador rezago se expresa de la siguiente manera:

$$\Phi(L) y_t = (1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) y_t = \varepsilon_t \qquad (1.2)$$

Un proceso AR(p) es estacionario en covarianza si y sólo si, las inversas de las raíces del polinomio en el operador rezago asociado al proceso autoregresivo ϕ(L), están dentro del círculo unitario. Si el proceso es estacionario en covarianza se puede escribir de la siguiente manera:

$$y_t = \frac{1}{\Phi(L)}\varepsilon_t \tag{1.3}$$

Una vez que el proceso cumple la condición de estacionalidad en la covarianza se puede afirmar que en la función de autocorrelación AR(p) decae en la mayoría de las formas de manera gradual de acuerdo al orden de la que es dependiente. Lo anterior simplemente significa que el modelo AR tendrá valores estadísticamente diferentes de cero en función a p innovaciones.

1.3 Modelos de medias móviles MA

Los procesos de medias móviles de orden finito también tiene una motivación directa: el hecho de que toda la variación en la serie de tiempo está explicada de manera importante por procesos estocásticos, por esto se puede estimar un modelo en el que se tomen en cuenta las series en el tiempo como rezagos atribuidos a choques actuales y pasados.

La forma como se expresa el proceso general de medias móviles de orden finito q, MA(q) es la siguiente:

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q} = \Theta(L)\varepsilon_t \tag{1.4}$$

En donde:

$$\Theta(L)\varepsilon_t = 1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q \tag{1.5}$$

Donde ε_t es un ruido blanco con varianza σ^2 y media cero. El proceso MA(q) toma en cuenta como parámetros las pautas pasadas más importantes, las cuales son valiosas

para ajustar mejor los pronósticos. Este proceso de orden q es invertible si cumple la mima condición de la raíz que el proceso de medias móviles, es importante mencionar esta relación ya que será importante cuando se junten estas dos teorías.

El polinomio es el operador rezago del MA(q) tiene q raíces; cuando q>1 surge la posibilidad de que haya raíces complejas. La condición de invertibilidad del proceso MA(q) es que las inversas de cada una de las raíces debe estar dentro del círculo unitario, en cuyo caso se tiene la representación autoregresiva convergente.

$$\frac{1}{\Theta(L)\varepsilon_t} y_t = \varepsilon_t \tag{1.6}$$

Al contario del proceso autoregresivo, la función de autocorrelación decae muy lentamente a cero. En contraste la función de autocorrelación parcial aplicada al modelo MA(q) decae de forma gradual de acuerdo al orden de la que es dependiente.

1.4 Modelos autoregresivo y medias móviles ARMA

Los modelos ARMA conjugan los Modelos AR y a los Modelos MA en una única expresión. Por tanto, la variable queda explicada en función de los valores tomados por la variable en períodos anteriores, y los errores pasados de la estimación. Una expresión general de un modelo ARMA (p, q) está dada por:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \tag{1.7}$$

Esta ecuación se puede resumir como,

$$\Phi(L) y_t = \Theta(L) \varepsilon_t \tag{1.8}$$

Donde el proceso (ε_t) es un ruido blanco con varianza σ^2 y media cero. Si las inversas de todas las raíces de $\phi(L)$ están dentro del círculo unitario, el proceso es estacionario en covarianza, y tiene una representación de medias móviles infinita convergente.

$$y_t = \frac{\Theta(L)}{\Phi(L)}\varepsilon_t \tag{1.9}$$

Si las inversas de todas las raíces de Θ(L) están dentro del círculo unitario, el proceso es invertible, y tiene una representación autoregresiva infinita convergente.

$$\frac{\Theta(L)}{\Phi(L)}y_t = \varepsilon_t \tag{1.10}$$

En la función ARMA no se puede determinar que se corta la función de autocorrelación y función de autocorrelación parcial en cierto orden sino que decae de forma gradual dependiendo el proceso.

1.5 Modelo ARIMA

Los modelos de series de tiempo analizados se basan en el supuesto de que las series de tiempo consideradas son estacionarias. En pocas palabras, la media y la varianza de una serie de tiempo débilmente estacionaria son constantes y su covarianza es invariante en el tiempo. Pero se sabe que muchas series de tiempo económicas son no estacionarias, es decir, son integradas.

Sin embargo, si una serie de tiempo es integrada sus primeras diferencias son estacionarias.

Por consiguiente, si se debe diferenciar una serie de tiempo d veces para hacerla estacionaria y luego aplicarle el modelo ARMA(p,q), se dice que la serie de tiempo original es ARIMA(p, d, q), es decir, es una serie de tiempo autoregresiva integrada de promedios móviles, donde p denota el número de términos autoregresivo, d el número de veces que la serie debe diferenciarse para hacerse estacionaria y q el número de términos de promedios móviles. (Gujarati & Down C, 2010).

El objetivo de la metodología Box – Jenkins es identificar y estimar un modelo estadístico que puede ser interpretado como generador de la información de la muestra. En este sentido, si el modelo estimado es usado para la predicción debe suponerse que las características de la serie son constantes en el tiempo, especialmente para los periodos

futuros. Por lo tanto, la predicción se efectúa sobre una base válida considerando que el modelo es estacionario o estable (Rosales), el método tiene 4 pasos:

1. Identificación. Consiste en transformar los datos, si es necesario, para que la hipótesis de estacionariedad sea adecuada, y en elegir los órdenes p,q

2. Estimación. Consiste en estimar el modelo ARMA(p, q)

3. Diagnosis. Consiste en comprobar que las propiedades empíricas corresponden a las hipótesis del modelo

4. Predicción. Utilizar el modelo para predecir

1.6 Identificación

El procedimiento para la identificación es una metodología más analítica basada en resultados, ya que se procede a identificar el número de coeficientes autoregresivos, errores de pronóstico pasados, además de determinar el número de diferencias que hagan que la serie no pierda su esencia. Es crucial este paso porque tomar muy pocos datos podría seleccionar información importante que puede no ser tomada en cuenta. Si se toman muchos datos podría hacer que el cálculo sea erróneo debida a información que no es relevante para la variable de estudio. En este proceso es inevitable pensar que las predicciones dadas por el modelo llevan consigo errores, la ventaja de la técnica estadística con que se hacen, es saber el nivel de confianza que se tiene.

Esta fase consiste en detectar el tipo de proceso estocástico que ha generado los datos. Esto significa encontrar los valores adecuados de p, d y q del modelo ARIMA. Las herramientas fundamentales en la identificación son la autocorrelación muestral y la autocorrelación parcial muestral. (Box y Jenkins , 1970)

Es importante tener en cuenta que antes de usar los criterios de identificación FAC y FAP se debe lograr una serie estacionaria. Para ello, se efectúan las pruebas de estacionariedad a la serie original. En caso de que esta no sea estacionaria, la variable puede diferenciarse d veces hasta que ésta sea estacionaria. Mediante este procedimiento se identifica el orden de integración d del modelo ARIMA.

1.6.1 Tendencia en varianza

Con el fin de volver estacionaria una serie, lo primero que se debe verificar es si es necesario eliminar la tendencia que esta tiene lo cual se puede realizar de varias formas. En general, se podrá obtener la serie transformada mediante el método de Box- Cox, el cual se refiere a elevar la serie a una potencia que estabilice la varianza. Este cálculo es más fácil determinarlo mediante software estadísticos en el que se obtiene un intervalo de potencia en el cual si el uno pertenece, no es necesario transformar la serie.

Si la serie no tiene tendencia en varianza, pero si en media, las desviaciones típicas o los rangos se mantendrán en un nivel similar al aumentar los valores de las medias calculadas, es decir, se observarán datos de forma horizontal.

1.6.2 Tendencia en media y ciclo

Una serie generalmente presenta tendencia, es importante determinar si la tiene, ya que se necesita eliminar para poder modelar correctamente los datos es importante mencionar que aunque se elimine, se agrega al final del modelado y no modifica el significado de los datos. El método de diferenciación transforma a la serie en estacionaria. Se obtiene restando del valor más reciente anterior, y se utiliza esta diferencia como una nueva serie. Cuando se realiza una diferenciación se analiza la gráfica para ver si sigue existiendo una tendencia.

Ya que se ha determinado la transformación que estabilice la varianza, el siguiente paso es estabilizar la media mediante el uso del operador diferencia $\nabla^k = (1 - B)^k$ y el operador de diferencia estacional $\nabla_k = 1 - B^k$ que se utiliza para estabilizar la tendencia de la serie y las tendencias en amplitud de ciclos de periodo de k unidades temporales también llamada la componente estacional.

Si en la serie existe una tendencia en la componente estacional de periodo de k unidades, este ciclo se elimina aplicando una o varias veces el operador$\nabla_k = 1 - B^k$.

La elección del número de veces d que hay que aplicar el operador retraso se decide graficando las series, junto con su respectiva FAC hasta conseguir eliminar la tendencia

en media. Análogamente se selecciona el valor k de diferenciación en el caso de una serie con tendencia en la componente estacional. [1]

1.6.3 Forma de las correlaciones en series estacionales

Para identificar el modelo, se sigue la metodología Box Jenkins usando los coeficientes de autocorrelación. Ese modelo autorregresivo identificado por medio de esta metodología explica las series de tiempo pausadas a través de ellas mismas rezagadas en el tiempo. El tiempo autorregresivo se refiere a que el pronóstico del valor de la serie se explica por el valor de la serie en tiempos pasados. En el proceso autorregresivo de orden p la observación actual es generada por un promedio ponderado de observaciones p periodos atrás.

Promedios móviles se refiere a los errores que se cometieron en los pronósticos pasados. Es una de las bases del modelo ARIMA. Contabilizando los modelos ARIMA puede alcanzar los cambios repentinos de la serie.

En el modelo de promedios móviles de orden q cada observación es generada por el promedio ponderada de perturbaciones aleatorias q periodos atrás.

A continuación, se presentan los patrones teóricos de la FAC y FAP según el tipo de modelo con una serie estacionaria, los cuales son útiles en la identificación de p y q del modelo ARIMA.

Tabla 1.1. *Comportamiento típico de los modelos AR, MA, ARMA*

MODELO	PATRON TIPICO FAC	PATRON TIPICO FACP
AR(p)	Decrecimiento rápido de tipo geométrico puro, geométrico con alternación de signos,	Picos grandes en los p rezagos o corta abruptamente a partir del rezago p, es decir la FAP

[1] Es muy importante no sobrediferenciar ya que podría afectar severamente la información de la serie.

	sinusoidal o mezcla de ambos tipos.	se anula luego del rezago p.
MA(q)	Picos grandes en los q rezagos o corta abruptamente a partir del rezago q, es decir la FAP se anula luego del rezago q.	Decrecimiento rápido de tipo exponencial y /o sinusoidal
ARMA(p,q)	Decrecimiento exponencial	Decrecimiento exponencial

Universidad de los Andes. Metodología Box-Jenkins. Pág. 2.

Puesto que en la práctica no se observan la FAC y la FAP teóricas, se usan las FAC y FAP estimadas, las cuales presentan cierto error estadístico. Lo que se busca es encontrar la mayor exactitud entre la FAC y FAP teóricas y estimadas, en tanto que la identificación del modelo ARIMA requiere de habilidad, la cual se obtiene con la práctica.

1.7 Estimación

En esta etapa se estiman los coeficientes de los términos autorregresivos y de media móvil incluidos en el modelo, cuyo número de rezagos p y q ya han sido identificados en la etapa anterior. Algunas veces la estimación se efectúa por mínimos cuadrados lineales, pero en otras se recurre a la estimación no lineal de los parámetros. Este último procedimiento utiliza un algoritmo para minimizar la suma de los cuadrados de los residuos, comenzando con algún valor inicial de los parámetros del modelo. En general el algoritmo busca si otro vector de parámetros mejora el valor de la función objetivo, produciendo iteraciones sucesivas hasta alcanzar la convergencia. Los paquetes estadísticos efectúan este procedimiento a través de rutinas de computador en las que se tienen definidos los parámetros iniciales, así como los criterios de convergencia. Teóricamente el método de mínimos cuadrados ordinarios en la medida que las muestras sean grandes posee propiedades asintóticas, esto quiere decir que se generan estimadores asintóticamente consistentes y convergen a una distribución normal, por lo que las pruebas hipótesis convencionales sobre los parámetros del modelo serán válidas.

La estimación del modelo ARMA(p,q) se efectúa para la serie que se ha comprobado es estacionaria. En la práctica los modelos más comunes son los autoregresivos. Hamilton cita que muchas de las series económicas se pueden representar por medio de un modelo AR(1). Sin embargo, de acuerdo con el teorema de descomposición de Wold, el modelo ARMA debería ser la primera opción, teniendo en cuenta que la inclusión de términos adicionales MA puede mejorar las propiedades estadísticas de la estimación. Los modelos MA son poco comunes y en la práctica a todos los modelos se les incorpora la constante o intercepto.

El modelo Box-Jenkins no solo tiene que ser estacionario, también tiene que ser invertible. Invertible reciente significa que las observaciones son más pesadas que las

observaciones más remotas; Los parámetros utilizados en el modelo disminuyen desde las observaciones más recientes hasta las observaciones posteriores. Los valores de t y los valores p aproximados prueban la siguiente hipótesis (O'Conell, Bowerman, & Koehler, 2005). Sea un parámetro particular en un modelo Box-Jenkins.

$$H_0 : \theta = 0 \; versus \; H_a : \theta \neq 0$$

Se puede rechazar la hipótesis nula, si y sólo si se cumple cualquiera de la siguiente condición:[2]

$$p - value < \alpha$$

Tenga en cuenta que si la hipótesis nula se rechaza en un nivel significativo menor, más fuerte es la evidencia de que el parámetro es importante para el modelo. Si uno de los parámetros es mayor al nivel de significancia es necesario que se quite para volver correr el modelo hasta que todos los parámetros sean significativos.

1.8 Diagnósis

1.8.1 Análisis residual

En esta etapa el objetivo es determinar si el modelo ARMA es adecuado para la serie correspondiente. Esto se comprueba verificando las siguientes hipótesis de los residuales:

Tabla 1.2.- *Hipótesis de los residuales y forma de comprobarla*

HIPOTESIS	VERIFICACION
Cov($\varepsilon_t, \varepsilon_{t-1}$)=0	FAC{ε_t}
Var(ε_t) es constante	FAC{$\varepsilon_t{}^2$} o gráfica de serie de tiempo de ε_t
ε_t~ Normal (0,1)	Test de normalidad (Jarque-Bera) o Gráfico de papel de probabilidad

[2] No es la única forma de calcular de evaluar los parámetros pero ésta es la más sencilla con buenos fundamentos matemáticos.

Fuente: Elaboración Propia

1.8.2 Selección de Modelos

Una vez que se han determinado el número de p y q, el siguiente paso es estimar el modelo. Los cálculos para tener los coeficientes del modelo son demasiado repetitivos para realizarlos manualmente, por eso es mucho mejor utilizar un software para ahorrar tiempo.

Es necesario cumplir las siguientes condiciones para poder decir que el modelo estimado es lo suficientemente bueno para detener el proceso de estimación:

- El cambio de los coeficientes de modelo a modelo debe ser pequeño
- La suma de los errores al cuadrado no cambia demasiado de un modelo a otro.
- Establecer un número límite de veces que los coeficientes serán revisados. Esto asegurará, que si por alguna razón la estimación continúa, puede ser detenida y se obtengan algunos resultados.

Es común que existan varios modelos para la misma serie. Existen supuestos que no tienen que ver con el análisis residual y necesitan ser verificados, estos son:

El modelo es considerado parsimonioso. En sí lo que la parsimonia implica es que no se puede reducir el número de parámetros involucrados, ya que todos son necesarios para explicar el comportamiento del fenómeno y no pueden ser considerados como iguales a cero. Lo que se hace es generar intervalos del 95% de confianza para los parámetros y verificar si el cero es un posible valor del parámetro.

Si el cero es un posible valor del parámetro, entonces se debe de cancelar y estimar el modelo sin él.

El modelo es admisible. La verificación para los modelos puede efectuarse fácilmente por simple inspección de que los parámetros estimados se encuentran dentro de las regiones admisibles.

1.9 Predicción

1.9.1 Pronóstico de series estacionarias

Sea {W_t} una serie de tiempo estacionaria con media cero, obtenida a partir de una serie {Z_t} con N observaciones. Como $W_t = \bigtriangledown^d T(Z_t)$ para algún valor d y para una cierta transformación T del tipo Box-Cox. Supóngase además que {W_t} admite una representación $W_t = \theta(L)a_t$ para el cual existe un modelo ARMA equivalente, modelo que se desea utilizar en la obtención de pronósticos de la serie. En particular si a partir el origen T, se desea pronosticar a la observaciónW_{t+h}, un pronóstico cualquiera d esta observación que se obtenga mediante una combinación lineal de los valores de la serie { W_t}, y en consecuencia de los errores a_t será denotado por $W_t(h)$ mientras que el pronóstico óptimo será escrito como$W_t^*(h)$.

El criterio que se usará para determinar la optimalidad del pronóstico será el error cuadrático medio mínimo, es decir,

$$W_t^*(h) = arg_{W_t} min E_T[W_{t+h} - W_t[h]]^2 \tag{1.11}$$

En donde E_T denota a la esperanza condicional, dada toda la información hasta el momento T, así:

$$E_T[W_{t+h} - W_t[h]]^2 = E\left\{[W_{t+h} - W_t(h)]^2 Z_T Z_{T-1}\right\} \tag{1.12}$$

Ya que $W_t(h)$ y por consiguiente $W_t^*(h)$ debe satisfacer.

$$W_t^*(h) = C_h a_T + C_{h+1} a_{T-1} + ... = \sum_{j=h}^{\infty} C_j a_{T+h-j} \tag{1.13}$$

El problema de obtener $W_t^*(h)$ se traduce en hallar los valores $C_h, C_{h+1}, \cdots$ de manera que se satisfaga la primera ecuación mostrada arriba. Con este fin la observación W_{T+h} se define como:

$$W_{T+h} = -\sum_{j=0}^{\infty} \theta_j a_{T+h-j} = -\sum_{j=0}^{h-1} \theta_j a_{T+h-j} - \sum_{j=h}^{\infty} \theta_j a_{T+h-j} \qquad (1.14)$$

con $\theta_0 = -1$

Donde la primer suma en la segunda igualdad corresponde a la información desconocida al tiempo T (abarca observaciones desde T+1 hasta T+h) mientras que la segunda consta de información conocida al tiempo T (desde menos infinito hasta T). De la segunda y tercera ecuación se obtiene:

$$W_{T+h} - W_t^*(h) = -\sum_{j=0}^{h-1} \theta_j a_{T+h-j} - \sum_{j=h}^{\infty} (C_j + \theta_j) a_{T+h-j} \qquad (1.15)$$

Entonces:

$$E_T[W_{T+h} - W_t^*(h)]^2 = \sum \theta^2 \sigma_a^2 + \sum (C_j + \theta_j)^2 \sigma_a^2 \qquad (1.16)$$

Donde el mínimo se alcanza cuando $C_j = -\theta_j$ para j=h,h+1,… por lo que el valor mínimo es:

$$E_T[W_{T+h} - W_T^*(h)]^2 = \sum \theta_j^2 \sigma_j^2 \qquad (1.17)$$

Y

$$W_T^*(h) = -\theta_h a_T - \theta_{h+1} a_{T-1} - \ldots = \sum \theta_j a_{T+h-1} \qquad (1.18)$$

$$E_T = \begin{cases} a_{T+h-j \to j \geq h} \\ 0 \to j < h \end{cases}$$

Por lo tanto de la tercera ecuación se obtiene:

$$E_T(W_{T+h}) = -E_T\left[\sum_{j=0}^{h-1}\theta_j a_{T+h-j}\right] - E_T\left[\sum_{j=0}^{\infty}\theta_j a_{T+h-j}\right] = -E_T\left[\sum_{j=h}^{\infty}\theta_j a_{T+h-j}\right] \qquad (1.19)$$

Por lo que $E_T(W_{T+h})$ proporciona el pronóstico con error cuadrático medio mínimo, es decir, $W_T^*(h) = E_T(W_{T+h})$.

Así el error del pronóstico con origen en T viene dado por :

$$e_t(h) = W_{T+h} - W_t^*(h) = -\sum_{j=0}^{h-1}\theta_j a_{T+h-j} \qquad (1.20)$$

Por lo cual:

$$E[e_t(h)] = 0 \text{ y } Var_T[e_t(h)] = \sum \theta_j^2 \sigma_a^2 \qquad (1.21)$$

De donde se concluye que los pronósticos $W_t^*(h)$ son insesgados y además como:

$$Var_T[e_t(h)] - Var_T[e_t(h-1)] = \theta_{h-1}^2 \sigma_a^2 \geq 0 \text{ para } h \geq 1 \qquad (1.22)$$

Entonces al emplear pronósticos óptimos, mientras más alejados se deseen los pronósticos (mayor sea h) mayor será a varianza (menor precisión) del mismo.

Ya que se desean obtener pronósticos de la serie $\{W_t\}$ considérese conocido el modelo $\phi(L)W_t = \theta(L)a_t$. Donde entonces:

$$W_t^*(h) = E_T[W_{T+h}] = E_T[\phi_1 W_{T+h-1} + \phi_2 W_{T+h-2} + ... + \phi_p W_{T+h} + a_{T+h} - \theta_1 a_{T+b-1} - ... - \theta_q a_{T+b-q}]$$

$$= \phi_1 E_T[W_{T+h-1}] + \phi_2 E_T[W_{T+h-2}] + ... + \phi_p E_T[W_{T+h}] + E_T[a_{T+h}] - \theta_1 E_T[a_{T+b-1}] - ... - \theta_q E_T[a_{T+b-q}] \quad (1.23)$$

Donde:

$$E_T(W_{T+h-j}) = \begin{cases} W_{T+h-j} \\ W_t^*(h-j) \end{cases} \begin{matrix} j \geq 1 \\ j \lhd h \end{matrix} \quad (1.24)$$

$$E_T(a_{T+h-j}) = \begin{cases} W_{T+h-j} - W^*_{T+h-j-1} \\ 0 \end{cases} \begin{matrix} j \geq h \\ j \lhd h \end{matrix} \quad (1.25)$$

1.9.2 Pronóstico de series no estacionarias

En la práctica es muy poco común observar series que originalmente sean estacionarias por lo cual es necesario generalizar lo visto anteriormente. Entonces, suponga que la estacionariedad de la serie original {Z_T} se cancela aproximadamente al determinar la transformación $T(Z_T)$ y aplicarle un número adecuado de diferencias, es decir, considerar $W_T = \bigtriangledown^d T(Z_T)$.

Para simplificar la exposición suponga que el grado de diferenciación requerido para cancelar la no estacionariedad homogénea es d=1 los pronósticos óptimos $W_T^*(h)$ podrán obtenerse de acuerdo con la teoría de la sección anterior y así los pronósticos óptimos de la serie {$T(Z_T)$} se obtienen simplemente de la relación:

$$E_T(W_{T+h}) = E_T[T(Z_{T+h})] - E_T[T(Z_{T+h-1})] = T^*(Z_T)(h) - E_T[T(Z_{T+h+1})] \quad (1.26)$$

Así entonces:

$$T^*(Z_t)(h) = \begin{cases} T(Z_T) + W_T(1) \;\; h = 2 \\ T^*(Z_t)(h-1) + W_t(h) \;\; h \geq 2 \end{cases} \quad (1.27)$$

1.9.3 Intervalos de confianza

Para hallar los intervalos de confianza para los pronósticos se hace uso de los resultados que se obtuvieron en la onceava ecuación. La cual aunada al supuesto de que $a_t \sim N(0, \sigma_a^2)$ para toda t implica que $e_t(h)Z_T, Z_{T-1}, ..., N \sim N(0, Var_T(e_T(h)))$.

Por lo que los límites al 100*$(1-\alpha)$% de confianza para $T(Z_{T+h})$.

$$T^*(Z_{T+1})(h) = T^*(Z_T)(h+1) - \theta_h a_{t-1} \text{ h>=1} \quad (1.28)$$

Esta expresión sirve de base para actualizar los pronósticos una vez que un dato más de la serie ha sido observado. Para hacer más explícita la utilidad de la ecuación anterior, supóngase que se han pronosticado H valores de la serie $T(Z_t)$ a partir del origen N, o sea, se han calculado los valores:

$$T^*(Z_N)(1), T^*(Z_N)(2), ..., T^*(Z_N)(H) \quad (1.29)$$

Supóngase además que el nuevo dato ya se observó, entonces, con h=1 se tiene:

$$a_{N+1} = T(Z_{N+1}) - T^*(Z_N)(1) \quad (1.30)$$

Así entonces estas ecuaciones permiten calcular los siguientes pronósticos con origen N+1:

$$T^*(Z_{N+1})(1) = T^*(Z_T)(2) - \theta_1 a_{N+1}$$

$$T^*(Z_{N+1})(2) = T^*(Z_T)(3) - \theta_2 a_{N+1}$$

.

.

.

$$T^*(Z_{N+1})(H-1) = T^*(Z_T)(H) - \theta_{H-1}a_{N+1}$$

Con lo anterior es posible calcular los pronósticos $T^*(Z_N)(h)$ junto con sus respectivos intervalos de confianza y, en su caso actualizarlos cuando se obtengan nuevas observaciones$Z_{N+1}, Z_{N+2}, \cdots$.

Por lo que los límites al 100*$(1-\alpha)$% de confianza para $T(Z_{t+H})$ condicionados al conocimiento de las observaciones $Z_T, Z_{T-1}, \cdots$, son

$$T^*(Z_T)(h) \pm Z_{\alpha_2}\left[\sum_{j=0}^{h-1}\theta_j^2\right]\sigma_a^2 \qquad (1.31)$$

El intervalo se obtiene al sustituir los parámetros por sus respectivas estimaciones.

1.9.4 Actualización del Pronóstico

En la práctica, las series de tiempo con las que se trabajan aumentan su número conforme el tiempo pasa, esto tiene efectos en particular sobre los pronósticos, puesto que el pronóstico $T^*(Z)(I)$ pide validez al conocerse la observación $T(Z_{T+1})$ de igual manera todos los $T^*(Z)(I)$ serán de poco utilidad si no se toman en cuenta las nuevas observaciones para actualizarlos. Con este fin, se puede utilizar la relación equivalente a la novena ecuación vista.

$$E_T[T(Z_{T+h})] = -\sum_{j=h}^{\infty}\theta_j a_{T+j-j} \qquad (1.32)$$

Para obtener

$$\Rightarrow T^*(Z_{T+1})(h) = T^*(Z_T)(h+1) - \theta_h a_{T+1} \Rightarrow h \geq 1 \qquad (1.33)$$

Expresión que sirve de base para actualizar los pronósticos una vez que un dato más de la serie ha sido observado. Para hacer más explícita la utilidad de la ecuación anterior, suponga que se han pronosticado H valores de la serie $T(Z_t)$ a partir del origen N, o sea, se han calculado los valores $T^*(Z_N)(1), T^*(Z_N)(2), ..., T^*(Z_N)(H)$; además suponga que el nuevo dato $T(Z_{N+1})$ ya se observó, entonces con h=1 se tiene:

$$\Rightarrow a_{N+1} = T(Z_{N+1}) - T^*(Z_N)(1) \tag{1.34}$$

Así la ecuación permite ahora calcular los siguientes pronósticos con origen en N+1:

$$T^{*+} = (Z_{N+1})(1) = T^*(Z_T)(2) - \theta_1 a_{N+1} T^{*+} = (Z_{N+1})(2) = T^*(Z_T)(3) - \theta_2 a_{N+1}$$

.

.

.

$$T^{*+} = (Z_{N+1})(H-1) = T^*(Z_T)(H) - \theta_{H-1} a_{N+1}$$

Con lo anterior es posible calcular los pronósticos $T^*(Z_N)(h)$ junto con sus respectivos intervalos de confianza y en su caso actualizarlos cuando se obtengan nuevas observaciones $Z_{N+1}, Z_{N+2}, \cdots$.

1.9.5 Pronósticos de la serie original

Este es uno de los tópicos más importantes, ya que brindará los datos pronosticados reales. En la mayoría de los casos, los pronósticos serán requeridos para la serie observada {Z_t} y no para la transformada $T(Z_t)$. En primer instancia lo más lógico es simplemente aplicar la transformada inversa para obtener la serie original.

Sn embargo, las propiedades óptimas del pronóstico $T^*(Z_t)(h)$ no se preservan necesariamente su la transformación T es no-lineal y se usa a:

$$T^{-1}[T^*(Z_t)(h)] \tag{1.35}$$

Como pronóstico de Z_{T+h}.

Como ya se ha visto anteriormente, una transformación usada con frecuencia es la potencia, ya sea en su versión original o en la normalizante. Para ambas transformaciones, es posible realizar el cálculo de un factor que permita corregir aproximadamente en seso que introduce la aplicación de la transformación inversa, para regresar a la escala original dicho factor puede estimarse mediante:

$$e_{T,\lambda}(h) = \begin{cases} \left\{ \frac{1}{2} + \sqrt{1 - 2\lambda(\lambda-1)[1+\lambda T^*(Z_t)(h)]^{-2} Var_T^{*[e_T(h)_2]}} \right\}^{1/\lambda} & si\ \lambda \leq -1\ y\ \lambda \neq 0 \\ exp\left\{ Var_T^{*[e_T(h)_2]} \right\},\ si\ \lambda = 0 \end{cases}$$

En el caso de la normalizante de Box- Cox y en el caso de la transformación potencia en si versión original se usa

$$e_{T,\lambda}(h) = \begin{cases} \left\{ \frac{1}{2} + \sqrt{1 - \frac{2(\lambda-1)}{\lambda}[1+\lambda T^*(Z_t)(h)]^{-2} Var_T^{*[e_T(h)_2]}} \right\}^{1/\lambda} & si\ \lambda \leq -1\ y\ \lambda \neq 0 \\ exp\left\{ Var_T^{*[e_T(h)_2]} \right\},\ si\ \lambda = 0 \end{cases}$$

Con estos factores, el pronóstico insesgado de Z_{t+h} será aproximadamente:

$$Z_T^*(h) = E_T[T^{-1}[T^*(Z_T(h))]] \approx T^{-1}[T^*(Z_T)(h)]\widehat{e_{T,\lambda}}(h) \qquad (1.36)$$

De igual manera como se corrigen los pronósticos, también se pueden corregir por sesgo los intervalos de confianza, de hecho se sugiere utilizar como intervalo del (1-alpha)*100% de confianza a:

$$T^{-1}[T^*(Z_T)(h)]\widehat{e}_{T,\lambda}(h) \pm z_{\frac{\alpha}{2}} \sqrt{Var_T^{\frac{*[e_T(h)]}{2}}} \qquad (1.37)$$

El cual incluye la corrección por sesgo.

La fase final del proceso de modelado de las series modelo ARIMA es aplicar la ecuación. Otra vez esta tarea matemática un software es el que mejor lo hace. Una característica importante del modelo ARIMA es poder determinar la probabilidad de acierto del pronóstico.

1.10 Variables Cualitativas o Dummy

Muchas veces es necesario enfrentarse a un problema donde lo importante es predecir valores cuyas variables de las que dependen son cualitativas. Ejemplos de variables cualitativas son: sexo del empleado, estado civil, jerarquía del empleado, etc. Estas variables a las cuales se les llama “dummy” o binarias se les considera en el modelo y se evalúa como explican la variable dependiente mediante la pendiente de las líneas que se generan.

Si se considera un modelo de regresión con una sola variable cualitativa A y una variable cuantitativa X. Es decir,

$$y = \beta_0 + \beta_1 x + \beta_2 A + \varepsilon \tag{1.38}$$

Si se consideran los casos:

Si A=0

Entonces,

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1.39}$$

Si A=1

Entonces,

$$y = (\beta_0 + \beta_2) + \beta_1 X + \varepsilon \tag{1.40}$$

El valor estimado de β_2 representa el cambio promedio en la variable de respuesta al cambiar el valor de la variable binaria.

Si se desea comparar las pendientes de las línea de regresión de los dos grupos se puede usar una prueba de t similar a la prueba de comparación de dos medias y asumiendo que hay homogeneidad de varianza.

Cuando la hipótesis nula no es rechazada se concluye que la pendiente de regresión de ambos grupos son iguales. Si no hubiera igualdad de varianza de los dos grupos, habría

que usar una prueba de t aproximada similar al problema de Behrens-Fisher. Aquí se usa una t con grados de libertad aproximados.

Al introducir las variables cualitativas al modelo, no solamente viene sustentado por el hecho de que se aproxima mejor sino que también tiene interpretaciones muy específicas dependiendo del tipo de variable cualitativa que se introduce.

1.11 Modelo ARIMA Modificado

Como se explica al inicio del capítulo III del presente trabajo (Modelos AR, MA y ARIMA para series de tiempo), la base fundamental para el trabajo sobre la series de tiempo es la regresión. Es decir, se puede determinar que el valor de una variable a través del tiempo no solamente se determina por el tiempo en sí, sino que además tendría factores categóricos que están dados y no varían en forma de choques aleatorios. Las variables cualitativas son factores que influyen en una variable sin estar determinados estocásticamente.

Es decir, se tiene determinado que existe un modelo ARIMA del que se puede extraer información para generar un pronóstico más o menos confiable, pero es necesario recalcar que tiende a ser muy volátil a medida que se aleja del punto focal. Además no toma en cuenta factores dados, y se enfoca totalmente en la aleatoriedad de los datos.

Es aquí donde pueden entrar estas variables cualitativas sumadas al modelo ARIMA, el cual está probado que es confiable. Esto se realiza sobre las bases de regresión.

Se tomará el modelo estimado por la metodología Box-Jenkins y a estos parámetros, que dependen de los valores de la variable en los tiempos anteriores y de los errores de estimación, se les sumarán las variables cualitativas para realizar la regresión. Este modelo tomará parámetros con variables estocásticas y determinísticas.

El modelo teórico está planteado y a simple vista parecería fácil realizarlo, pero si se analiza a profundidad. Es un proceso difícil conocer los residuales del modelo proyectado durante ciertos periodos inciertos de valores reales para con ello tener la información

suficiente para realizar la regresión con el vector correspondiente, ya que el modelo tendría que correrse para cada periodo lo cual haría hasta cierto punto trivial el objetivo del pronóstico.

Durante el proceso de selección del mejor modelo ARIMA para la serie, uno de los supuestos que se deben cumplir en un buen modelo ARIMA es que los residuales tendrán distribución normal. Es decir, al tener el ajuste del modelo y los datos reales hasta el último dato, se tiene el vector de residuales pero además de eso es necesario tener los residuales de los valores pronosticados con el ARIMA óptimo.

El vector de residuales del pronóstico no se puede calcular de la misma manera para los valores que ya se tienen que para los que se pronostican. Sin embargo, al saber que se distribuyen de manera normal se conocen los parámetros básicos para simularlo. Es decir, se tienen que simular determinados datos igual al horizonte de pronóstico. El vector original de los residuales sumado a la simulación de éstos para el pronóstico, generarán los datos suficientes y confiables para tomar esos datos como un regresor confiable.

En la regresión, la variable independiente es sobre la que se desea hacer el pronóstico y las variables independientes, serían los datos de los retrasos correspondientes a los factores AR y el vector de residuales con datos reales y simulados correspondientes a los factores MA del modelo ARIMA óptimo. Sumado a estos regresores, se toman como variables independientes las variables dummy o cualitativas correspondientes a la característica de temporalidad de la serie, es decir, si son meses, años, semanas, días.

Es importante notar que la regresión será tomada sobre un vector de residuales que ha sido simulado. Sin embargo, una simulación muchas veces no da la proximidad a la distribución deseada. Es por ello que es necesario realizar determinadas iteraciones que den la certeza de la exactitud deseada.

Como ya se conoce, el proceso de regresión es iterativo en el que se tiene que estar evaluando qué valores independientes son significativas para el mejor modelo[3]. No

[3] Generalmente al 95% de confianza

obstante, hay que tener en cuenta que se tienen que evaluar muchos modelos de regresión y sus parámetros dependiendo de los datos simulados, en este caso de los residuales. Generalmente, las simulaciones son procesos iterativos de gran requerimiento computacional por lo tanto evaluar cada modelo de regresión iterativamente es un proceso complejo.

Con ayuda de software computacional es posible evaluar automáticamente el mejor modelo de regresión correspondiente a las determinadas variables independientes dadas. Este proceso se tiene que hacer con cada simulación sobre los residuales.

El modelo, aunque en la simulación se tengan los parámetros como se debe distribuir, tendrá diferentes parámetros. Es decir, no en todos los modelos de regresión óptimos correspondientes para cada simulación se tendrán los mismos parámetros. Para ello se tendrá que elegir el que tiene mayor frecuencia, lo cual asegurará que es el mejor para con este pronosticar.

Sin embargo, es necesario tomar en cuenta que la metodología Box Jenkins destaca que ciertos términos autoregresivos y de promedios móviles son significativos para la serie de tiempo. Por lo cual hay evidencia empírica y teórica de la significancia de estos regresores para la variable dependiente. Es por esto que, el modelo de regresión óptimo para describir el comportamiento de la variable dependiente debe contener entre sus regresores a los parámetros determinados por el ARIMA además de los posibles regresores cualitativos a evaluar.

Este en su forma más estricta no sería una serie de tiempo, sin embargo, sí lo es. Implícitamente los regresores AR y MA tienen dependencia del tiempo. Utilizando un modelo de regresión se utilizan los valores de las variables independientes para pronosticar sobre la serie sobre la cual se realizó el modelo. El pronóstico depende de un vector de residuales el cual no se tiene. Sin embargo, al igual que se realiza para determinar la regresión óptima, es necesario simular este vector, pero de nueva cuenta, no es difícil debido a la existencia de los parámetros para realizarlo.

De la misma manera que se hizo para determinar la regresión óptima, se debe tomar la media de los pronósticos para tener la certeza del mismo. Esta herramienta permite

también calcular la distribución del pronóstico para cada tiempo, y con ello generar un intervalo de confianza del mismo.

La metodología Box-Jenkins genera un modelo dependiente de determinados factores AR y MA .En general, da una muy buena estimación para aquellos pronósticos de los que toma valores reales. Sin embargo, a medida que depende de valores generados por el modelo su pronóstico se comporta más erróneo. Es por ello que el pronóstico óptimo necesita tomar a los valores que predice el ARIMA pero solamente para el número máximo de AR o MA del modelo. Después de dicho valor, es necesario, tomar el pronóstico generado por el modelo que también toma en cuenta las variables cualitativas.

II. Contextualización del problema

Para la economía de cualquier país, es importante tener en consideración que existen complejas estructuras del trabajo debido a la diversidad de sociedades. La discordancia entre las clasificaciones utilizadas en todas las regiones es otro tema con el que hay que lidiar.

Para tratar esta diversidad de rubros, se han creado clasificaciones típicas que componen actividades con personal capacitado, técnico, cualificado, no cualificado, etc. Además de clasificaciones aceptadas por la macroeconomía como el sector primario, secundario y terciario.

La información estadística laboral es esencial para conocer las características y evolución del mercado laboral además de diseñar políticas públicas orientadas a mejorar las oportunidades y condiciones de empleo. Derivado de la necesidad de una medición confiable, en México, una de las mediciones más utilizadas es la llamada PEA (Población Económicamente Activa) con base en esta, se pueden tomar clasificaciones acerca del desempleo con encuestas como la ENOE (Encuesta Nacional de Ocupación y Empleo) donde se obtiene información estadística sobre las características ocupacionales de la población, así como otras variables demográficas y económicas. La cobertura estatal y de ciudades ha sido cubierta de manera muy efectiva por la ENOE debido a la mayor demanda en las ciudades y a nivel municipal (Secretaría del Trabajo y Previsión Social, 2008). Estas formas de medición han sido consecuencia de un gran trabajo realizado durante los últimos tiempos. Se puede decir que son los suficientemente confiables para entender el contexto ocupacional en México, además de ser indicadores comparables con otros países.

2.1 La política de empleo en el contexto de México

Durante la década de los setenta, México tenía una situación muy parecida a la que tenían todos los países latinoamericanos con características preocupantes como el excesivo endeudamiento externo, déficit de balance en los gobiernos con las empresas paraestatales además de estructuras industriales no competitivas etc. Debido al

agotamiento de los recursos en empresas del Estado, se derivó una tenencia hacia el aislamiento del Estado en la vida económica, por lo cual se realizaron transferencias de recursos desde el sector estatal hacia el sector privado fundamentales para las empresas. Debido a este comportamiento al final de los años setenta, durante la década de los ochenta se vio reducido el PIB además de que se incrementó el abono de recursos al exterior debido al incremento de la deuda externa, esto tenía implícito el aumento del desempleo además de la reducción crónica del ingreso real de los trabajadores y programas sociales. La propuesta económica que el país planteó a la problemática del desempleo se fundó en la teoría neoclásica con la idea que el sistema económico tiene una tendencia natural al equilibrio siempre que no existan fuerzas perturbadoras que obstaculicen su funcionamiento (Barcelata Chávez).

En ese tiempo el gobierno mexicano vio modificado su sistema de pagos en gran parte por el Fondo Monetario Internacional y para adaptarse a los requerimientos del capital internacional, por lo tanto, se desmanteló el proteccionismo estatal insertándose así a la economía neoliberal asumiendo los requerimientos necesarios.

Como menciona (Harvey, 2005) el Estado neoliberal debe favorecer fuertes derechos de propiedad privada individual y las instituciones de libre comercio. El marco legal está determinado por las obligaciones contractuales libremente negociadas entre los sujetos en el mercado. Este nuevo orden trajo como consecuencia la reducción del gasto público en programas sociales, además de a venta de muchas de las empresas paraestatales con el objetivo de hacer más eficiente la economía en base a la racionalización. Todas estas acciones trajeron consecuencias preocupantes como el incremento de los despidos masivos no solamente en las empresas que eran en su totalidad propiedad de gobierno, también en las que sus operaciones dependían del sector público lo que amplió la brecha entre el nivel de empleo, que el aparato productivo o de servicios puede contratar, y el nivel de oferta de mano de obra disponible. Es evidente que se limitaron la creación de nuevos empleos además de mermar en gran parte otros puestos de trabajo formales de la economía lo cual llevó a niveles muy altos la desocupación y subempleo.

La situación mencionada anteriormente aunada a las tendencias demográficas recientes como el incremento de la PEA hace que exista una sobreoferta de la fuerza laboral en la que el mismo mecanismo desvaloriza la fuerza de trabajo. Estos graves problemas tienen indicios claro como el incremento de los índices de pobreza, marginación, altas tasas de migración y cierre de numerosas empresas.

En la década de los noventa donde ya había un fuerte proceso de globalización, un fuerte desajuste financiero desencadenó problemas en el PIB de México afectando sectores intensivos en mano de obra. Este fue uno de las muestras que la estabilidad del empleo y la generación de nuevos trabajos habían dejado de depender del gobierno y había pasado a manos de las empresas privadas lo que exigía acelerar la incorporación de nuevas tecnologías y formas de organización encaminadas al empleo más eficiente, proceso que sigue hasta la actualidad.

2.2 El desarrollo del empleo a lo largo del tiempo

El desempleo en sí, ha sido unos de los problemas sociales más graves en México. Esto afecta gravemente la economía del país debido a la fuga de población económicamente activa del país. Esta situación no solo afecta a la mano de obra sino al personal capacitado que también migra en busca de mejores oportunidades.

Existe un discordancia entre la tasa de crecimiento de la Población Económicamente Activa (PEA) y el ritmo de crecimiento del empleo remunerado. El número de empleos requeridos para cubrir las necesidades del incremento anual de la oferta laboral era de poco más de un millón cien mil (Censos Nacionales de Población y Vivienda, Instituto Nacional de Estadística, Geografía e Informática, México). Es evidente que existe un déficit en el empleo acumulado pero es difícil de precisar.

En la actualidad es necesario que la economía de México genere aproximadamente 1.2 millones de empleos debido a que esto sería suficiente para cubrir el aumento de la PEA pero desafortunadamente está lejos de lograrlo ni siquiera lo puede lograr debido al creciente flujo migratorio. Aunque el empleo creció entre 2004 y 2008 no lo hizo de manera suficiente para cubrir rezagos importantes. Otro de los problemas graves que se

puede observar es que los empleos formales parecen ser más informales. Ya que no se cuentan con las prestaciones sociales suficientes. Una de las características de inicios del siglo XXI es el crecimiento de la economía informal sin protección social lo cual es generado por el incremento de los niveles de desempleo. Se observa, además, un amplio vacío para generar puestos de trabajo en sectores industriales. Según cifras oficiales actualmente hay cerca de 1.5 millones de desocupados más que en 2000. Además, en este mismo periodo la tasa de presión general (porcentaje que representa la población desocupada más la ocupada que busca trabajo, respecto a la población económicamente activa) creció en al menos 2 veces en 22 de las 32 entidades de la República. (Ruíz Nápoles & Ordaz Díaz).

Según datos de la Organización de Cooperación y Desarrollo Económico (OCDE) México es un país estable en cuento al panorama laboral pero estas estadísticas pueden ser engañosas debido a la falta de evaluación del empleo informal, una de las características preocupantes de la economía nacional. Según datos del Instituto Nacional de Estadística y Geografía (INEGI) en México laboran aproximadamente 29.3 millones de personas en la informalidad, lo cual quiere decir que seis de cada diez personas no tienen seguridad social.

La crisis del 2008 golpeó a México en uno de los peores momentos: cuando la pirámide demográfica se ensanchaba cada vez más en la parte media, es decir, afectó de manera fuerte a una mayor cantidad de personas. Esto junto con las reformas estructurales afecta severamente al empleo. Hay una gran cantidad de jóvenes que tratan de ingresar a la fuerza laboral y que no lo están haciendo debido al poco crecimiento del empleo formal y que ha obligado a la fuerza de trabajo a buscar salidas. Una de ellas son el nuevo fenómeno denominado Ni-Ni, simplemente se refiere a población joven que no cuenta con un trabajo y no estudia, lo que genera una situación preocupante. Como bien se especifica en (Situación del Empleo en México, 2013) en un principio, este fenómeno pasó desapercibido en las estadísticas, ya que esta población —que debería ser parte de la fuerza de trabajo de nuestro país- pasó a formar parte de la Población no Económicamente Activa (PEI), por lo que salió de la estadística que debería conformar el desempleo y al salir de la PEA, quedó encubierta. Los Ni-Ni son personas en edad

laboral que no están incapacitadas, no están jubiladas, no están a cargo de tareas del hogar y cuentan con un nivel de escolaridad superior al de las generaciones pasadas; sin embargo, no estudian ni buscan empleo. Muchos de ellos declaran estar disponibles para trabajar, pero no realizan una búsqueda activa de empleo por estar desalentados al haberlo intentado y no lograrlo.

Es importante tomar en cuenta que la educación es uno de los pilares fundamentales de la sociedad, ya que con una sociedad bien educada se generan las condiciones para tener una mejor economía en general, donde haya menos desigualdades. Es por todo esto, que es importante cuidar y proteger el mercado laboral de las personas que cuentan con un grado alto (o medianamente alto) de educación. Si este mercado está descuidado entonces existe el riesgo de que la sociedad se desincentive a educarse y esto traiga más problemas de los que ya hay.

Las condiciones para la inversión generadas por el gobierno han traído un impacto positivo sobre las tasas de desempleo pero es importante hacer notar que no se ha hecho un esfuerzo conjunto para que las personas con altos grados de educación encuentren trabajo de manera más fácil.

Es un problema importante debido a que en los últimos años se han hecho esfuerzos en la creación de preparatorias de gobierno además de universidades en todos los estados del país por lo cual se estarán integrando al mercado laboral muchos nuevos profesionistas y técnicos que no encontrarán donde ejercer su profesión.

Según la Secretaría de Educación Pública, proyectado a 2018, se pretenden tener 69 nuevas universidades cuyo objetivo principal de esto es ampliar la matrícula de estudiantes de educación media superior lo cual son aproximadamente poco menos de un millón de estudiantes más. (Diario Oficial de la Federación, 2013)

El presente trabajo llamado Pronóstico de la tasa de desocupación para personas con educación media superior y superior en México para 2018, plantea analizar los datos mensuales del indicador de desempleo brindado por el banco de información económica de INEGI y con ello tener estimadores de la tasa de desempleo calculados con un alto grado de confianza.

Al finalizar el trabajo se contará con los valores proyectados de todos los meses de 2018 lo cual es una base fundamental para que se pueda trabajar, teniéndolos se pueden contrastar las reformas que en su momento aplique el gobierno además de analizar si los proyectos de inversión disminuyen de manera considerable las tasas de desempleo estudiadas.

La tasa de desempleo es uno de los indicadores más importantes de una economía, inclusive algunos lo catalogan por encima del crecimiento del PIB debido a que deslumbra de una mejor manera el bienestar de las familias en la economía.

Como bien especifica Heath (2012) es muy importante ya que es uno de los primeros indicadores que se dan a conocer y con su divulgación empieza el reto de armar el rompecabezas de la marcha de la coyuntura. Es de las variables que más puede influenciar el comportamiento de los mercados si es que se aparta mucho de la expectativa.

En cambio, en México en general bajo el ojo público se le otorga poca importancia. Prácticamente no hay reacción de los mercados, aún en el caso de que la cifra reportada resulta ser muy diferente a la esperada. Únicamente los medios le dan un lugar en las primeras planas, pero siempre y cuando la tasa sea mayor al mes (o año) anterior. Esta importancia del desempleo toma un gran alcance al evaluar a las personas que son semiprofesionalizadas y profesionalizadas en una economía, ya que mide las tendencias de los alcances de las reformas además de medir si el mercado laboral tiene la capacidad suficiente para absorber a los técnicos y profesionistas.

Es por todo esto, que se refleja la importancia de tener un estimador confiable de la tasa de desempleo en personas con educación media superior y superior ya que no solamente da un panorama de la desocupación sino que también es un reflejo del mercado laboral y la economía nacional para el 2018.

2.3 Medición del desempleo en México

La fuente más confiable para conocer el porcentaje de desempleo en México es la Encuesta Nacional de Ocupación y Empleo (ENOE). La ENOE es la encuesta más levantada en México. Se empezó a utilizar en el 2005 y se vio caracterizada por un replanteamiento integral de los procesos de encuesta utilizados hasta el momento

además de cambiar los controles de calidad, comunicación y retroalimentación con los distintos operadores activos.

Utiliza métodos probabilísticos para la selección de la muestra aplicando la encuesta a los miembros del hogar seleccionado, utiliza un muestreo bietápico, estratificado y por conglomerados.

Como se especifica en INEGI (2017), la encuesta está diseñada para identificar sin confundir los conceptos de desocupación, subocupación e informalidad, así como también para tomar en cuenta y darles un lugar específico a aquéllas otras personas que no presionan activamente en el mercado laboral porque ellas mismas consideran que ya no tienen oportunidad alguna de competir en él (mujeres que por dedicarse al hogar no han acumulado experiencia laboral, personas maduras y de la tercera edad, etc.).

La encuesta tiene un enfoque global del empleo ya que no solamente le importa el desempleo en sí, sino que esa diseñada para tomar en cuenta aspectos como la marginación y la calidad del empleo. Por todo esto, se puede afirmar que es una fuente confiable de datos en la cual se tendrá la seguridad de que son correctos además de que el muestreo haya pasado por filtros de seguridad en la cual la información es eficaz.

2.4 Panorama del desempleo en México

En los último años debido al incremento en la cobertura de educación media superior y superior (EMS y S) no es dificil proponer que existe la incorporación de nuevos profesionistas al mercado laboral debido al aumento.Absorber los puestos de trabajo es un reto difícil para la economía de un país. Esta situación no es única de la oferta educativa superior, sino que se ha visto esta tendencia para las personas que han egresado de la eduación media superior. Como se muestra en (Panorama General de los Indicadores Laborales en México, 2017) , hasta el tercer trimestre del año 2016 se tenía un 5.3% de desocupación para personas con EMSy S, situación que ha ido en evidente crecimiento durante el último año.

De acuerdo con Burgos Flores & López Montes (2010) se puede determinar que, el problema más importante no es sólo el desempleo de profesionistas, sino las condiciones en las que llegan al mercado laboral, en el que no gran parte de estos egresados no

obtienen las remuneraciones adecuadas además de desempeñar puestos en los que no es necesario tener una gran cantidad de conocimientos; otro factor es cuando no tiene mucha coincidencia el trabajo desarrollado con la carrera estudiada o simplemente no aplica totalmente los conocimientos y habilidades adquiridos durante su educación. Este fenómeno de la carencia de empleo puede rechazar supuestos de la economía en la que se dice que las empresas buscan aprovechar los cnocimientos y habiilidades de los empleados para optimizar sus recursos.

Es importante tener una explicación no solamente causal de las situaciones sino que también analizar el desempleo acorde al tiempo y cómo tiene cambios naturales conforme avanza|. En general, se puede decir que una variable puede ser afectada por variables independientes, pero esto no siempre define que el modelo sea el mejor, en cambio, se puede decir que esta variable solamente depende del tiempo en sí para saber cuál será su comportamiento. Es por ello que la herramienta básica para pronosticar es una serie de tiempo.

Como mencionan Box Jenkins una serie de tiempo es el conjunto de datos que se registran a través del tiempo sobre el comportamiento de una variable de interés, generalmente y para una mejor estimación, los registros se realizan en periodos iguales de tiempo. Las series de tiempo resultan especialmente útiles cuando se requiere realizar un pronóstico sobre el comportamiento futuro que puede tener una variable determinada. Resulta útil el análisis de las series de tiempo que representan, bajo la hipótesis de que los factores que han influenciado su comportamiento en el paso, estarán presentes de manera similar en el futuro. De esta manera, el objetivo principal del conocimiento de las series de tiempo es la identificación de los factores que intervienen y la separación de cada uno de ellos, con el fin de pronosticar cuál será el comportamiento futuro.

Tomando entonces la serie de tiempo del desempleo como base, se comenzará el procedimiento y tratamiento de la información para poder pronosticarla para todo el año 2018 con un alto grado de confianza.

III. Adaptación y Aplicación del Modelo a los datos para el pronóstico

La utilidad de realizar un análisis de los datos mensuales del desempleo en México para personas con educación media superior y superior se fundamenta en la posibilidad del uso de esta información en la planeación económica y financiera de las empresas además de ser un indicador clave de cómo se encuentra la economía del país.

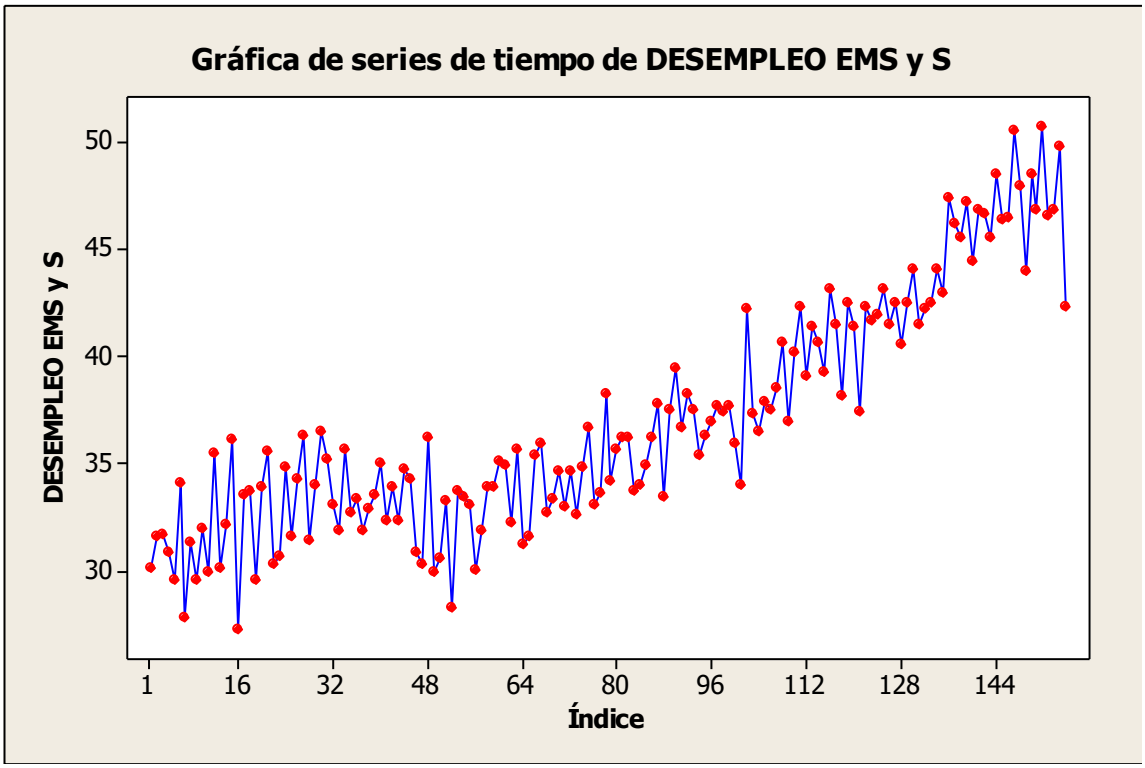
Con el fin de cumplir en su totalidad el objetivo general de la presente tesis, se analizan los datos históricos del porcentaje de población desocupada con educación media superior y superior ; porcentajes obtenidos de la Encuesta Nacional de Ocupación y Empleo (ENOE) la serie consta de 168 datos mensuales, desde Enero de 2004 hasta Diciembre de 2017, se tomó este periodo en el que hubo una fuerte crisis económica, fines de sexenio y cambio del partido gobernante en la presidencia de la república.

Los modelos ARIMA para estimar modelos con datos reales son una combinación de criterio y bases matemáticas, una combinación prudente de ellos acarreará por consecuencia, un modelo confiable con el cual se pueda pronosticar.

Los gráficos presentados en este capítulo se realizaron con la ayuda del software Minitab además de RStudio mediante los paquetes para series de tiempo.

Lo primero que se hace es graficar la serie de tiempo, esto no es más que poner en orden temporal los datos, desde el primer hasta el último. Una vez que se tienen graficados, es necesario analizarlos, primero de forma visual para poder determinar si existe una tendencia en la media además de la varianza de los datos. Después de analizarse visualmente, es necesario proceder con las herramientas que le darán sustento a lo ya observado.

Gráfico 3.1. *Comportamiento de la tasa de desocupación para personas con educación media superior y superior*



Fuente: Elaboración Propia con datos del BIE

En la serie se observa que tiene una tendencia, esto quiere decir que representa el comportamiento predominante de la serie, es el cambio de la media a lo largo de un periodo. Se puede observar que tiene una fuerte tendencia hacia arriba.

Con base a lo ya observado, se puede determinar que la serie no es estacionaria, es decir, a simple vista su media y su varianza no son constantes. Como ya se ha determinado en la primera parte de este trabajo, para poder aplicar la metodología Box-Jenkins de series de tiempo, es necesario que los datos temporales tengan la característica de media cero además de ser una banda de la misma anchura, es decir, que los datos tengan la misma varianza. A las series con estas características se les llaman series estocásticas.

3.1 Series Estocásticas

Una caminata aleatoria con deriva, es un modelo de tendencia, porque en promedio crece cada periodo el valor de la deriva. Así, el parámetro de la deriva desempeña el mismo papel que el de la pendiente en un modelo de tendencia lineal determinística.

A la caminata aleatoria con deriva se le llama modelo de tendencia estocástica, porque la tendencia está impulsada por choques estocásticos, en contraste con las tendencias deterministas.

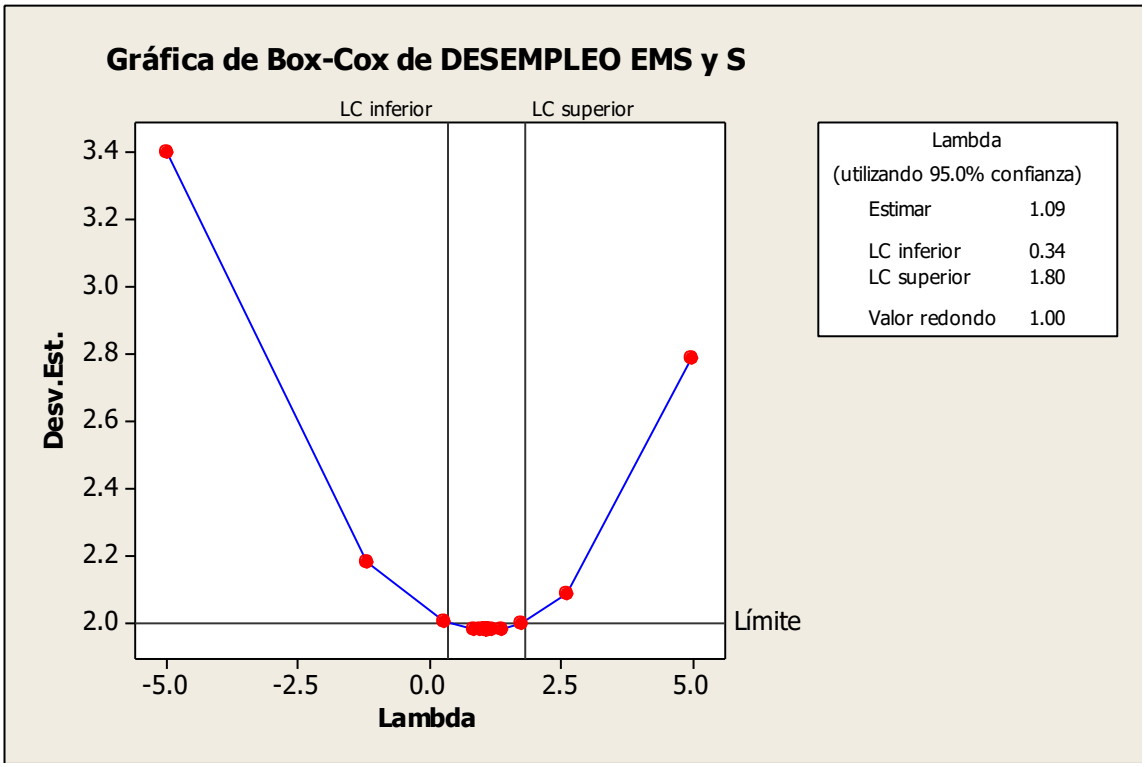
3.2 Estabilización de la varianza

El primer paso para tener la serie estocástica a partir de los datos originales es estabilizar la varianza de la serie, para ello es necesario utilizar la transformación de Box- Cox. Las transformaciones de Box y Cox son una familia de transformaciones potenciales usadas en estadística para corregir varianzas desiguales (para diferentes valores de la variable predictora). Esta transformación recibe el nombre de los estadísticos George E. P. Box y David Cox.

La transformación simplemente se refiere a elevar toda la serie a una potencia[4] que estabilice la varianza para los datos.

[4] En el caso de que la potencia que estabilice la varianza sea cero se utilizará el logaritmo natural

Gráfico 3.2. *Transformación de Box-Cox aplicada a la serie analizada*



Fuente: Elaboración Propia con datos del BIE

La familia de transformaciones potencia determinada por el análisis Box-Cox, evalúa una infinidad de potencias, es muy probable que no solamente una sea la que se pueda adaptar a los datos sino que exista un intervalo en el cual la varianza permanezca constante al elevar los datos a cualquiera de esas potencias.

Es por ello que al realizar el análisis de Box-Cox en Minitab, muestra el gráfico 3.2 se determina que hay dos valores que podrían ser óptimos para los datos analizados. Pero no solo eso, sino que muestra un intervalo en el cual, al utilizar cualquiera de los datos que estén dentro del intervalo, los datos analizados no perderían la característica de homocedasticidad.

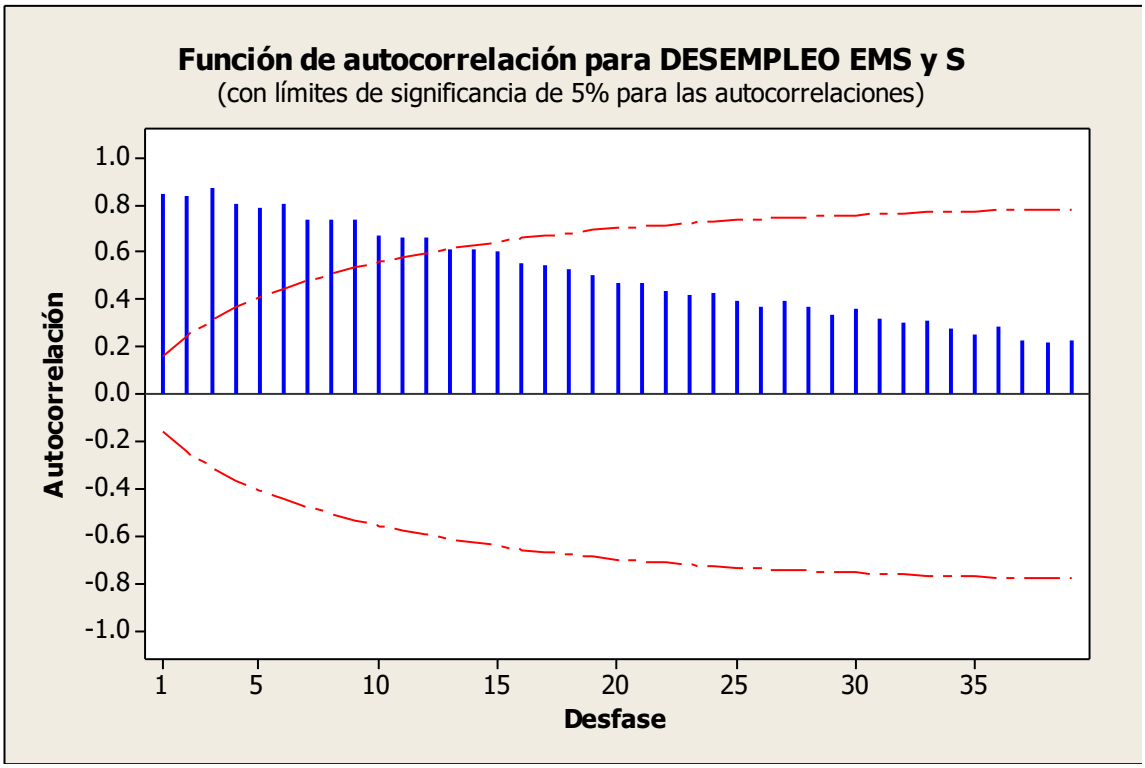
Como se puede observar 3.2 el uno pertenece al intervalo de confianza al 95%, con lo cual se puede determinar que lo datos que se están analizando, ya cuentan con la

característica de homocedasticidad. Una vez demostrada la varianza constante, es necesario analizar el correlograma con lo cual se podrá determinar si existe una tendencia en media.

3.3 Estabilización de la media

Una vez que existe la homocedasticidad en la serie de tiempo, es necesario observar el correlograma y el gráfico de serie tiempo transformado[5], con este se puede determinar si es necesario estabilizar la media.

Gráfico 3.3. *Comportamiento función de autocorrelación de serie de tiempo del desempleo para EMS y S*



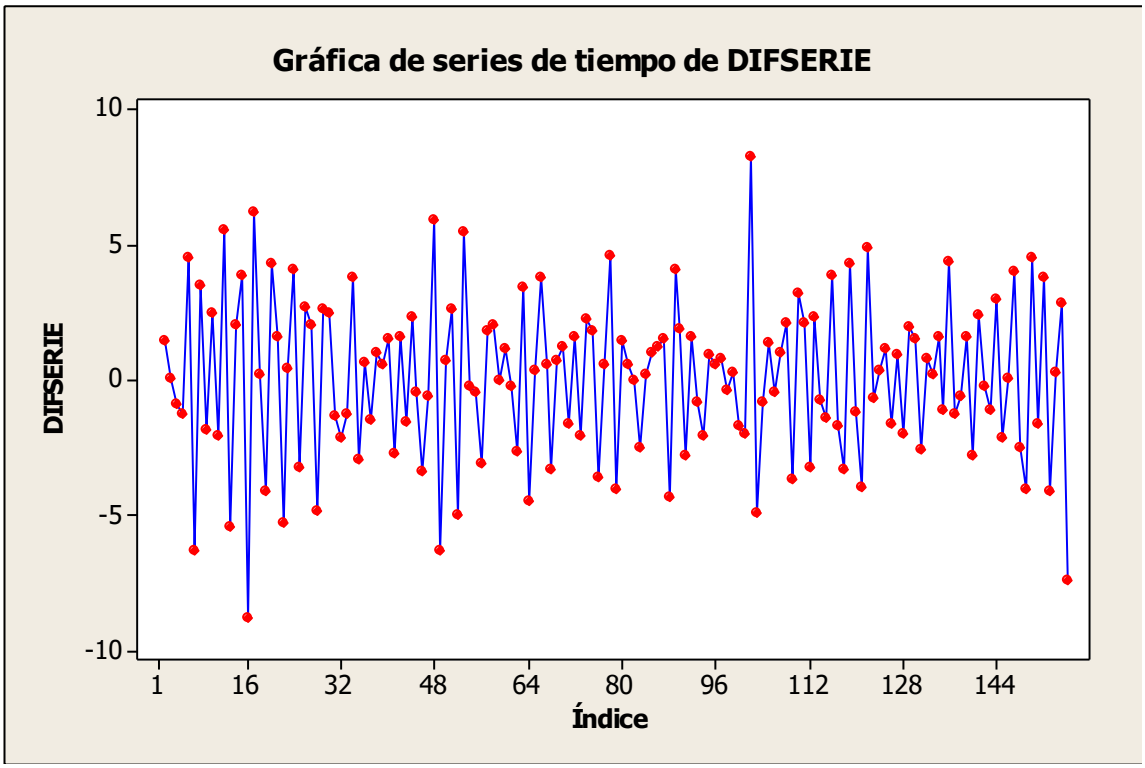
Fuente: Elaboración Propia con datos del BIE

[5] En este caso la serie transformada es igual a la original

Es muy evidente observando el correlograma, que estos no decaen exponencialmente a cero (estadísticamente) como se observa en el gráfico 3.3 lo cual, como ya se ha especificado anteriormente, muestra que los datos no se mantienen constantes a lo largo del tiempo, además en la serie de tiempo (gráfico 3.1) se ve una clara tendencia alcista durante todo el recorrido entonces es necesario utilizar una herramienta para que los datos tengan las características de una serie modelable. Para lograr este objetivo se utilizan las diferencias sobre la serie.

Como se observa en el gráfico 3.1, hay una tendencia a la alza. De esta manera no se pueden realizar los cálculos para determinar un modelo. Para corregir esto, se aplica la primera diferencia cuya serie de tiempo resultante se muestra en el gráfico 3.4.

Gráfico 3.4. *Serie de tiempo de la primera diferencia de la serie.*



Fuente: Elaboración Propia con datos del BIE

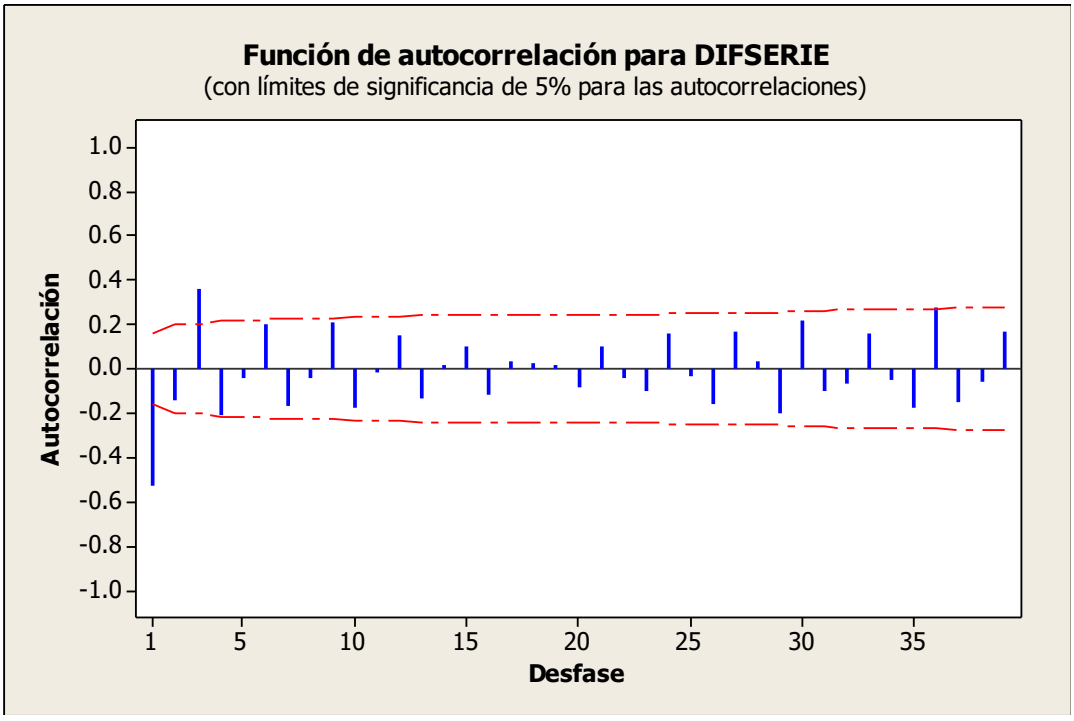
Se observa como aplicada la primera diferencia se tiene una banda del mismo ancho que además cuenta con media cero, entonces se puede determinar que la primer diferencia de la serie es un proceso aleatorio modelable.

El siguiente paso es analizar el gráfico de correlación y correlación parcial para determinar los términos autorregresivos y de promedios móviles para posteriormente estimar el modelo.

3.4 Análisis de la función de autocorrelación (FAC)

Una vez que se tiene una serie modelable, es necesario determinar los órdenes del modelo ARIMA. Para el componente de promedios móviles (MA) se utiliza la función de autocorrelación. Se centra la atención en los retrasos que son estadísticamente diferentes de cero, es decir, los valores que están fuera del intervalo de confianza mostrado en el gráfico siguiente.

Gráfico 3.6. *Función de Autocorrelación de la serie diferenciada.*



Fuente: Elaboración Propia con datos del BIE (INEGI,2017)

Función de autocorrelación: DIFSERIE

Desfase	ACF	T	LBQ
1	-0.522412	-6.50	43.13
2	-0.143179	-1.43	46.39
3	0.357656	3.53	66.87
4	-0.212970	-1.95	74.18
5	-0.041578	-0.37	74.46
6	0.204547	1.83	81.29
7	-0.163963	-1.44	85.71
8	-0.038657	-0.33	85.96
9	0.209265	1.81	93.26
10	-0.171589	-1.45	98.20
11	-0.017770	-0.15	98.25
12	0.149201	1.25	102.04
13	-0.131593	-1.09	105.01
14	0.017947	0.15	105.06
15	0.103614	0.85	106.93
16	-0.118894	-0.97	109.40
17	0.033849	0.27	109.61
18	0.021637	0.18	109.69
19	0.013463	0.11	109.72
20	-0.084105	-0.68	111.00
21	0.097729	0.79	112.73
22	-0.040449	-0.33	113.03
23	-0.096337	-0.78	114.74
24	0.160963	1.29	119.55
25	-0.035829	-0.28	119.79
26	-0.162486	-1.29	124.77
27	0.167348	1.31	130.10
28	0.032492	0.25	130.30
29	-0.197380	-1.53	137.83
30	0.215199	1.64	146.84
31	-0.099804	-0.75	148.80
32	-0.070912	-0.53	149.79
33	0.156682	1.17	154.69
34	-0.050541	-0.37	155.20
35	-0.174320	-1.29	161.36
36	0.278343	2.04	177.21
37	-0.149767	-1.07	181.83
38	-0.058596	-0.41	182.55
39	0.168669	1.19	188.52

Fuente: Elaboración Propia utilizando MINITAB con datos del BIE (INEGI,2017)

Es importante de decir que, como se busca el principio de parsimonia, se centra la atención en los valores fuera del intervalo de confianza pero que estén más cercanos a cero, es decir, se toma de izquierda a derecha el menor número de valores que sean estadísticamente diferentes de cero[6], en este caso se puede observar que el último valor

[6] Es muy común que en la práctica se tome como máximo valor de 5

que se sale del intervalo de confianza al 95% es tres por lo que se puede afirmar que el término de promedios móviles es tres[7].

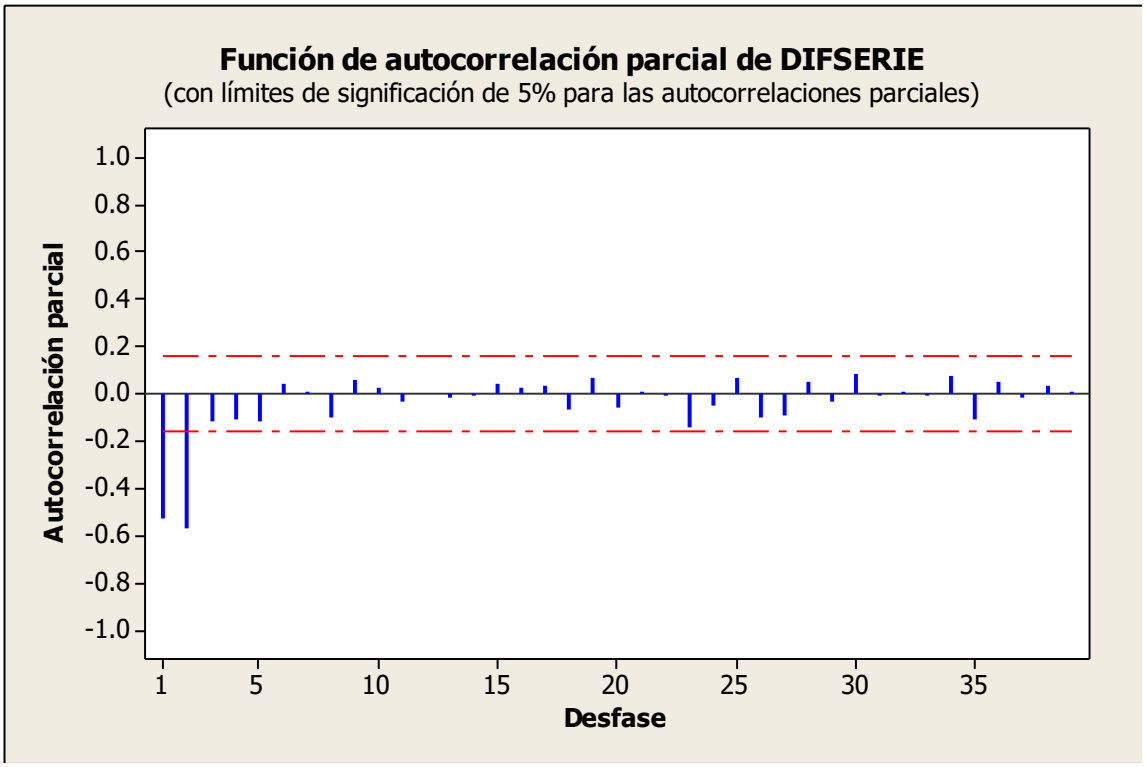
Este valor p no necesariamente será el que contendrá el modelo final, sino que es necesario hacer pruebas en conjunto con los términos autorregresivos para formular el modelo que se utilizará para el pronóstico. Para esta seria, el valor inicial con el que se empezarán a evaluar los modelos es 3, es decir, se tomarán 3 factores de promedios móviloes.

3.5 Análisis de la función de autocorrelación parcial (FACP)

Ya que se tiene determinado el orden de los promedios móviles (MA) es necesario analizar la función de autocorrelación parcial de la cual se extraerá información para el término autorregresivo(AR), para ello se grafica la función de autocorrelación parcial.

7 También se puede determinar mediante el uso del desfase y su correspondiente prueba T, en este caso también se determina que el valor es tres

Gráfico 3.7. *Función de Autocorrelación parcial de la serie diferenciada.*



Fuente: Elaboración Propia con datos del BIE

Función de autocorrelación parcial: DIFSERIE

Desfase	PACF	T		Desfase	PACF	T
1	-0.522412	-6.50		14	-0.009837	-0.12
2	-0.572276	-7.12		15	0.041887	0.52
3	-0.120483	-1.50		16	0.024785	0.31
4	-0.111212	-1.38		17	0.029680	0.37
5	-0.119350	-1.49		18	-0.069563	-0.87
6	0.039682	0.49		19	0.067148	0.84
7	0.008634	0.11		20	-0.056251	-0.70
8	-0.097129	-1.21		21	0.006160	0.08
9	0.060272	0.75		22	-0.010004	-0.12
10	0.026217	0.33		23	-0.144128	-1.79
11	-0.029408	-0.37		24	-0.048061	-0.60
12	0.000924	0.01		25	0.067678	0.84
13	-0.020142	-0.25		26	-0.100699	-1.25
				27	-0.094453	-1.18

28	0.052875	0.66	34	0.074936	0.93
29	-0.037594	-0.47	35	-0.106596	-1.33
30	0.086645	1.08	36	0.052152	0.65
31	-0.011509	-0.14	37	-0.013597	-0.17
32	0.006738	0.08	38	0.031758	0.40
33	-0.006623	-0.08	39	0.005167	0.06

Para determinar los términos autoregresivos de esta serie se utiliza el mismo criterio de selección utilizado en la función de autocorrelación. Se centra la atención en los retrasos que son estadísticamente diferentes de cero, es decir, los valores que están fuera del intervalo de confianza mostrado en el gráfico 3.7. En el caso de esta serie, se puede observar en los valores T de la prueba y en la función graficada que se tendrán dos términos autoregresivos en el modelo de prueba.

3.6 Análisis del modelo ARIMA

Una vez precisados los términos p,q correspondientes a promedios móviles MA y autoregresivos AR, es necesario evaluar qué tan bueno es el modelo ARIMA conjunto. Para ello se toman los términos de la sección 5.3 y 5.4 y se forma el modelo ARIMA correspondiente, entonces se evaluará el modelo ARIMA(2,1,3)[8].

Este proceso de asignación de parámetros en toma en base a las pruebas de las funciones de autocorrelación y tendencia. El primer parámetro referente al proceso autorregresivo, el segundo en base a la tendencia y el tercero en base a las medias móviles.

Por lo tanto se tiene el modelo siguiente:

Estimados finales de los parámetros

Tipo		Coef	SE Coef	T	P
AR	1	-0.8240	0.1651	-4.99	0.000
AR	2	-0.5805	0.1595	-3.64	0.000
MA	1	0.1697	0.1894	0.90	0.372
MA	2	0.1138	0.1990	0.57	0.568
MA	3	0.0967	0.1666	0.58	0.562

[8] Se toma el uno de la primer diferencia realizada a la serie

Se utilizará el valor p para evaluar cada uno de los parámetros tipo AR y MA. El valor p es una medida de la fuerza de la evidencia en sus datos en contra de H0. Por lo general, mientras más pequeño sea el valor p, más fuerte será la evidencia de la muestra para rechazar H0[9]. Más específicamente, el valor p es el menor valor de α que conduce al rechazo de H0. Para cualquier valor de α > valor p, usted no puede rechazar H0, y para cualquier valor de α $\leq$ valor p, usted rechaza H0. Tradicionalmente, el valor p se compara con valores de α menores que .05 o .01, dependiendo del campo de estudio en tes trabajo se trabajará a un nivel de confianza del 95% por lo que el valor de α tiene que ser menor o igual a .05.

Para el modelo ARIMA (2, 1, 3) es necesario anular el último término MA por lo que corresponde evaluar un ARIMA (2 ,1, 2).

Este es un proceso en el que se debe de evaluar cada modelo y con ello, todos los parámetros que contengan los correspondientes modelos. Utilizando este proceso reiterativo se llega al modelo final:

Modelo ARIMA (2, 1, 2)

Estimados finales de los parámetros

Tipo		Coef	SE Coef	T	P
AR	1	-0.7757	0.1783	-4.35	0.000
AR	2	-0.4867	0.0879	-5.54	0.000
MA	1	0.2334	0.1940	1.20	0.231
MA	2	0.1720	0.1682	1.02	0.308

Estadística Chi-cuadrada modificada de Box-Pierce (Ljung-Box)

Desfase	12	24	36	48
Chi-cuadrada	5.1	9.6	26.3	34.2
GL	8	20	32	44

[9] En este caso Ho es que el correspondiente parámetro AR o MA no debe estar en el modelo ARIMA final.

Valor P	0.751	0.974	0.751	0.856

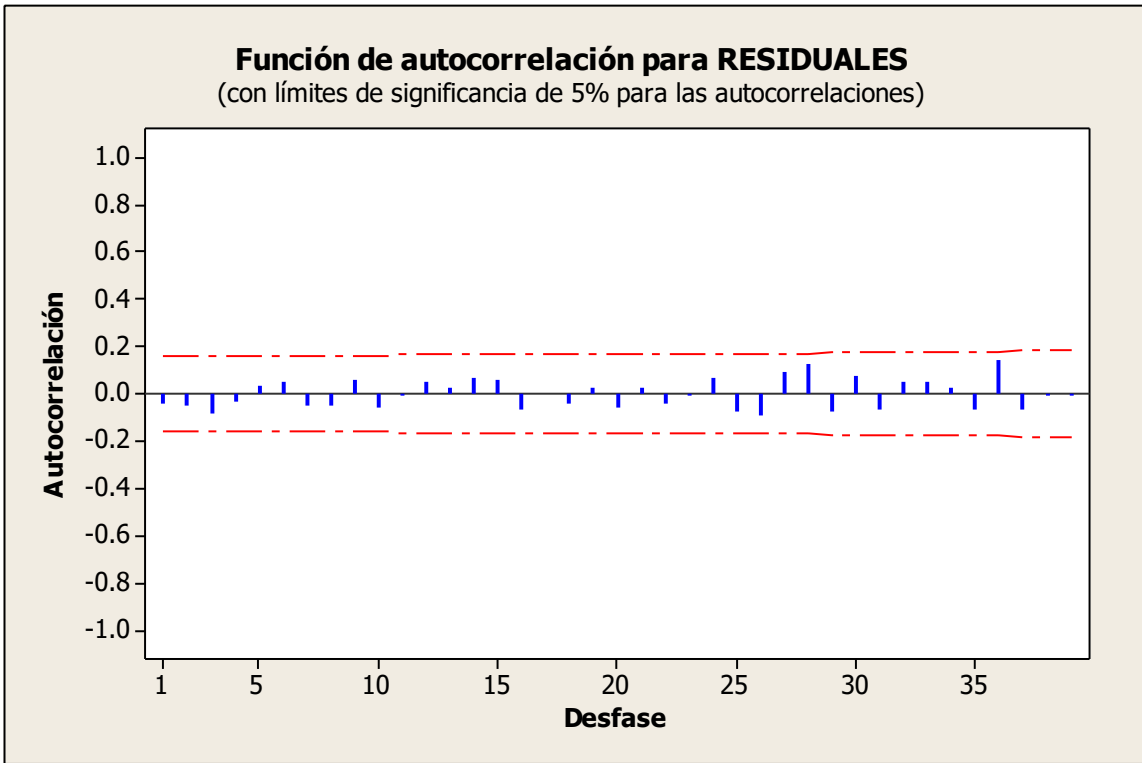
El modelo a simple vista pareciera no ser óptimo debido a las pruebas correspondientes a los términos MA pero al anular el último componente MA si bien se corrigen los valores p de las pruebas de los parámetros, el valor p de la prueba de Ljung-Box[10] muestra un comportamiento no deseado por lo que se determina mejor quedarse con el modelo ARIMA (2, 2, 1).

Una vez que se ha determinado cuál modelo contiene los términos AR y MA óptimos para la serie, es necesario evaluar los supuestos para comprobar si el modelo es apropiado.

El primer supuesto que se evaluará es la correlación existente entre los residuales, este análisis se basa habitualmente en los residuos que no deben estar correlacionados con el pasado: su correlograma no debe tener ninguna correlación significativamente distinta de cero además el estadístico de Ljung-Box en sus valores de prueba p deben ser mayores al nivel de confianza utilizado en el modelo.

[10] El estadístico q de Ljung-Box se utiliza para comprobar si una serie de observaciones en un período de tiempo específico son aleatorias e independientes. Si las observaciones no son independientes, una observación puede estar correlacionada con otra observación k unidades de tiempo después, una relación que se denomina autocorrelación. La autocorrelación puede reducir la exactitud de un modelo predictivo basado en el tiempo, como la gráfica de series de tiempo, y conducir a una interpretación errónea de los datos. Principalmente, se utiliza para evaluar los supuestos después de ajustar un modelo de series de tiempo, como ARIMA, para asegurar que los residuos sean independientes.

Gráfico 3.8. *Autocorrelación de los residuales.*



Fuente: Elaboración Propia con datos del BIE

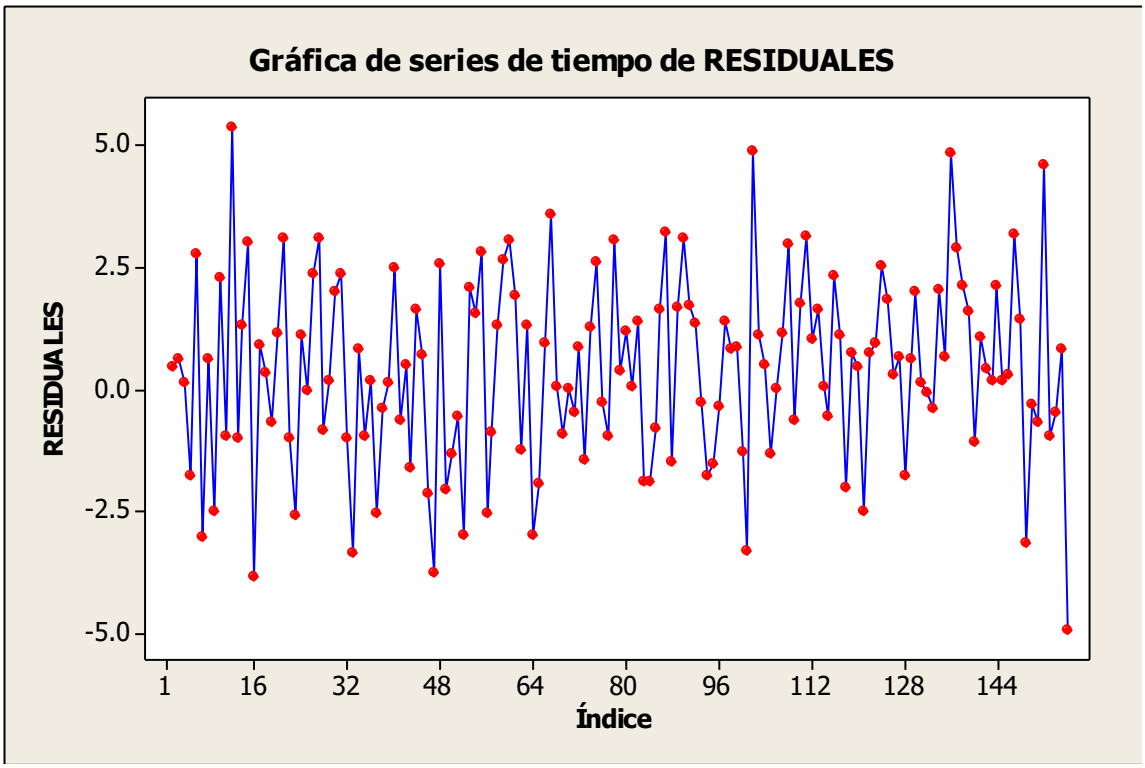
Estadística Chi-cuadrada modificada de Box-Pierce (Ljung-Box)

Desfase	12	24	36	48
Chi-cuadrada	5.1	9.6	26.3	34.2
GL	8	20	32	44
Valor P	0.751	0.974	0.751	0.856

Como se observa en el gráfico 3.8 además en el estadístico Ljung-Box se puede determinar que los residuos no tienen relación entre sí por lo que no brindan más información importante que pudiera ser útil.

Con el siguiente supuesto se trata de comprobar que los residuales tengan una media cero y se dispersen de una manera simular, para ello es necesario graficar la serie de tiempo de los residuos como se muestra en el gráfico 3.9:

Gráfico 3.9. *Serie de tiempo de los residuales*

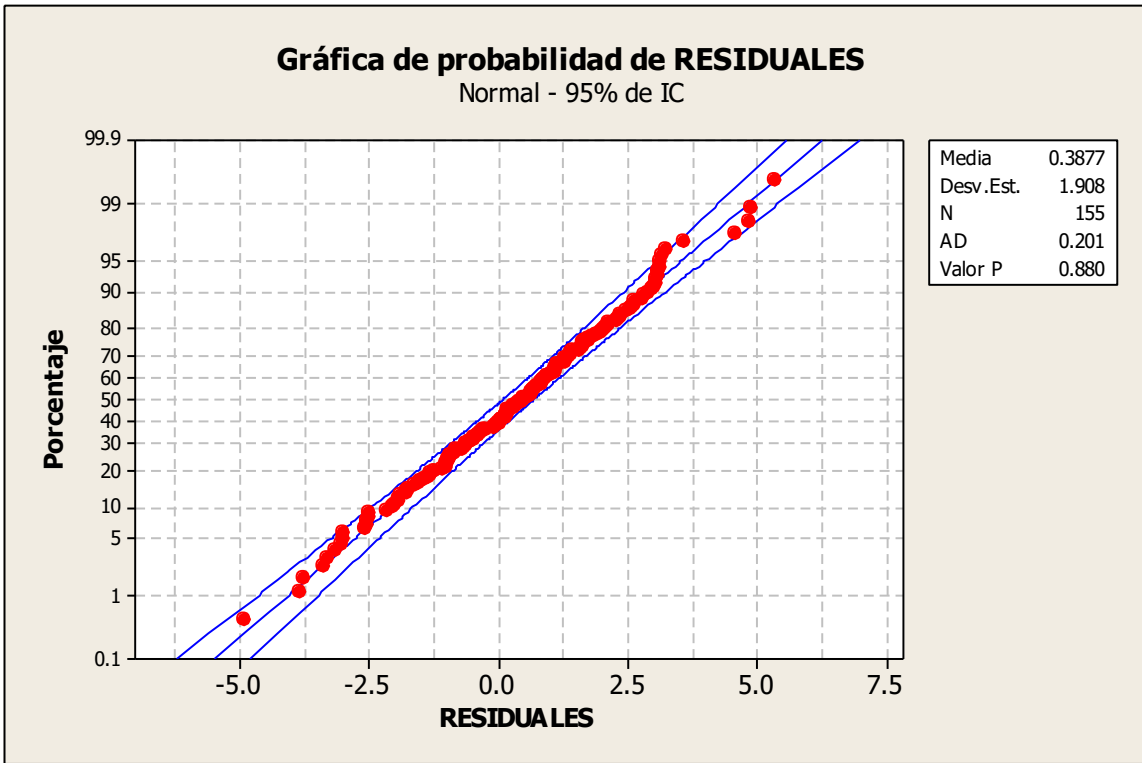


Fuente: Elaboración Propia con datos del BIE (INEGI ,2017)

En el grafico 3.9 se puede observar claramente que los residuos no tienen tendencia además que se desplazan en una banda del mismo ancho alrededor de cero por lo que se determina que cuentan con la característica de varianza constante.

El último supuesto a evaluar es si los residuales se distribuyen de manera normal, para ello se realizará el gráfico de papel de probabilidad en el que se podrá evaluar el valor p para comprobar si se cumple la hipótesis nula Ho mostrado en el gráfico 3.10 .

Gráfico 3.10. *Prueba de normalidad de los residuales*[11].



Fuente: Elaboración Propia con datos del BIE

En la gráfica de papel de probabilidad la hipótesis nula Ho dice que los datos cuentan con una distribución normal así que si el valor p es menor al nivel de significancia hay evidencia suficiente para rechazar Ho.

Utilizando un nivel de confianza de 95% se tiene un valor p esta prueba para los residuales de 0.880 hay una clara muestra de que no se puede rechazar Ho por lo que los residuales sí se distribuyen de manera normal (con media 0.3877 y desviación estándar de 1.908).

[11]MINITAB nombra la prueba de normalidad como gráfica de probabilidad

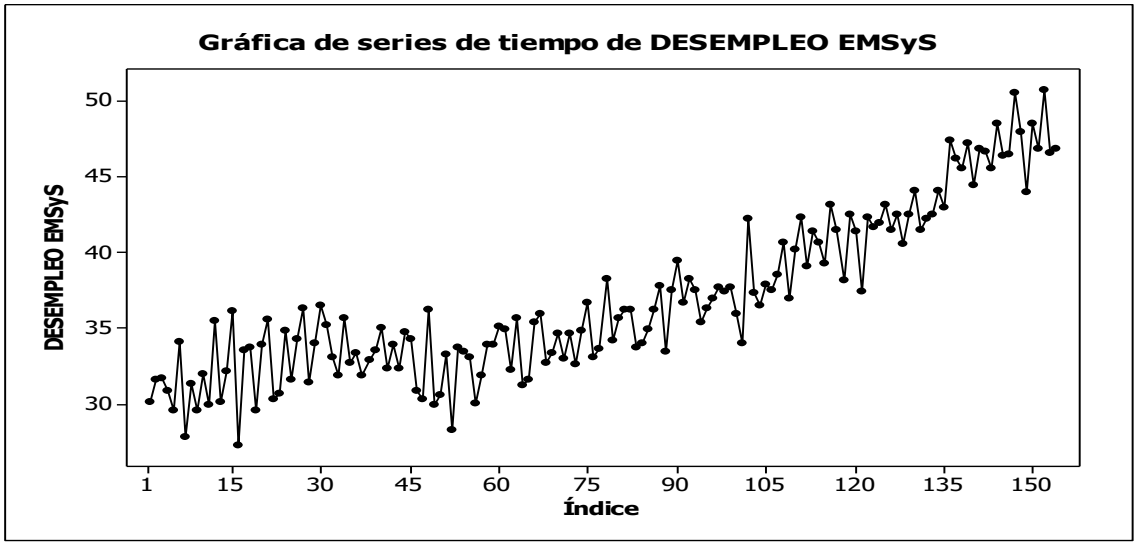
Una vez que se ha verificado que los residuales cumplen los supuestos es posible someter los datos a un pronóstico con la suficiente confianza para saber que se acercarán a los reales.

3.7 Pronóstico del modelo ARIMA

El objetivo de ese trabajo es ver cómo se comporta la tasa de desempleo de personas con educación media superior y superior para con esto analizar y utilizar esta herramienta para pronosticar.

El estudio se realizó con la tasa de desempleo de personas con educación media superior y superior mensual desde 2005 hasta 2017, información que se puede ver en el gráfico 3.11. Utilizando información oficial pubicada por el INEGI y proveniente de la base de datos recopilada por la ENOE.

Gráfico 3.11. *Serie de tiempo del desempleo para EMSyS.*



Fuente: Elaboración Propia con datos del BIE

El análisis de estos datos se realizó empleado la metodología específica de los modelos ARIMA. El modelo propuesto como proceso generador de los datos pasó de forma satisfactoria las pruebas de diagnóstico a las cuales fue sometido: normalidad, no correlación de los residuales de forma individual además de varianza constante de estos. Se identificó un proceso no estacionario ARIMA (2, 1,2).

Se puede observar que los datos tienen una media que desplaza a través del tiempo con tendencia a aumentar, sin varianza grande entre cada periodo.

Cuadro 3.1: *Parámetros estimados del modelo ARIMA (2,1,2)*

Parámetros	θ_1	θ_2	φ_1	φ_2
Estimación	-0.7757	-0.4867	0.2334	0.1720

Fuente: Elaboración Propia

El modelo puede ser usado para evaluar el comportamiento de la tasa de desocupación en personas con EMSyS. Esto se verificó empleando los pronósticos de un periodo por delante desde enero de 2017 hasta diciembre de 2017, haciendo una comparación con los valores observados. En la gráfica de la figura 3.12, se muestran estas predicciones con su respectivo intervalo de confianza al 95 por ciento, graficando también la información que se tiene en ese periodo. Si se llega a presentar un dato fuera de estos intervalos, es posible que sea un valor irregular, que debe ser sometido a revisión. Por lo cual, el modelo que mejor representa el comportamiento de los datos de acuerdo a la metodología Box-Jenkins es:

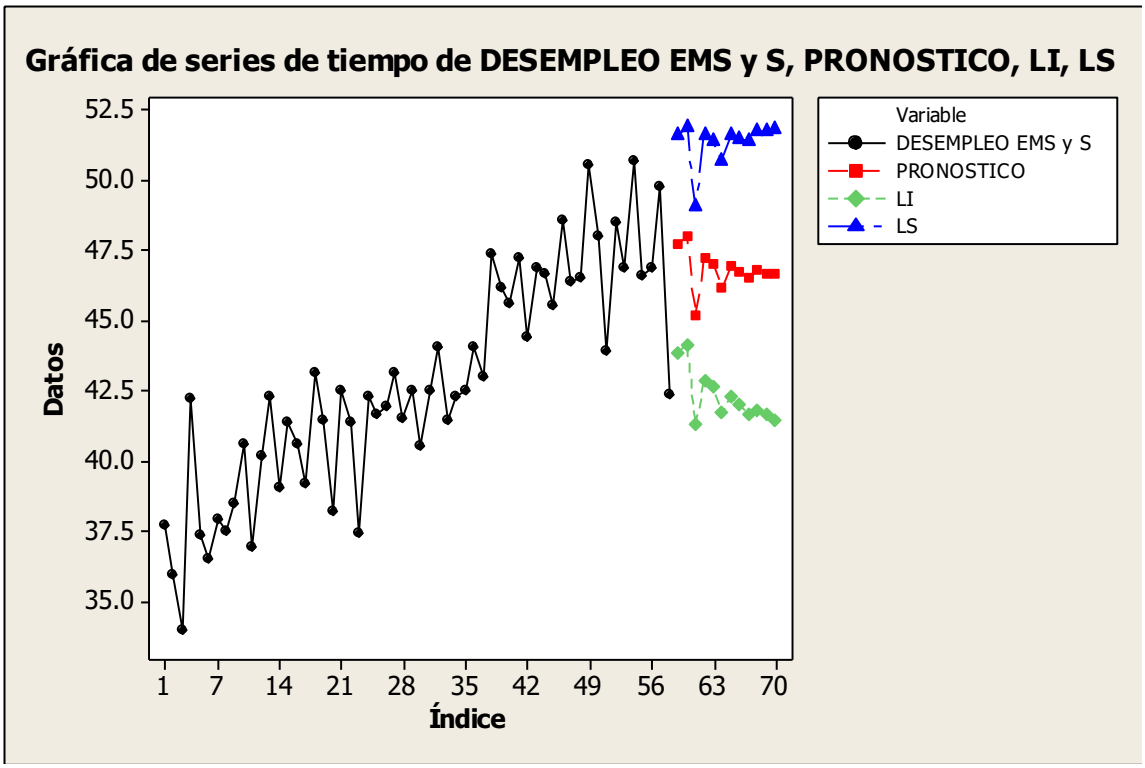
$$y_t = -0.7757y_{t-1} - 0.4867_{t-2} + 0.2334\varepsilon_{t-1} + 0.1720\varepsilon_{t-2} + \varepsilon_t \qquad (3.1)$$

De este modelo específico se puede deducir que el valor mensual del desempleo en personas con educación media superior y superior depende negativamente un 77% del valor del mes inmediato anterior y un 48% del valor del mes antepasado al que se desea

conocer, además de depender un 23% y un 17% del error del modelo del mes inmediato anterior y del mes antepasado al que se desea conocer respectivamente.

Lo que se puede exponer claramente es que el modelo depende fuertemente de dos valores inmediatos anteriores, en este caso de los dos meses más recientes por lo cual, se puede inferir que el pronóstico tendrá una inestabilidad clara debido a la autodependencia del modelo. Esto se puede observar en el gráfico 3.12.

Gráfico 3.12. *Pronóstico del desempleo en EMSyS.*



Fuente: Elaboración Propia

Para analizar la calidad de las predicciones del modelo se realizaron pronósticos dentro de un intervalo temporal para el cual se tienen datos. Partiendo del origen de diciembre del 2017 se estimaron los pronósticos de hasta 12 periodos hacia delante (gráfico 3.12).

En la variable que MINITAB llama límite, se representa la serie de tiempo en todos sus puntos incluyendo el pronóstico. En la variable llamada Datos, se muestra el valor de la serie para cada punto del tiempo.

Se observa que durante los primeros dos meses las predicciones son aceptables, pero como se observa a partir de marzo de 2018 los pronósticos no son tan buenos; lo anterior debido a la dependencia que tiene el modelo de los dos meses inmediatos anteriores por lo cual el modelo tendrá que tomar datos pronosticados por el mismo modelo, lo cual creará una estabilidad del pronóstico por lo tanto la varianza de las predicciones crece al aumentar la distancia al origen del pronóstico como se observa en los intervalos de confianza.

3.7.1 Propuesta para disminuir la varianza a través de variables cualitativas y simulación

Como se puede observar claramente en el gráfico 3.12. Existe un problema de varianza en el pronóstico, es decir, a medida que los datos se van alejando del punto focal, la línea de predicción tiende a ser horizontal, lo cual simplemente causaría grandes residuales en esos valores. Lo mismo se puede observar en los intervalos de confianza, simplemente tienden a abrirse lo cual generaría el nivel de confianza deseado pero no la exactitud del pronóstico requerida.

Lo más adecuado es generar variables artificiales correspondientes a cada mes del año, con la meta de cubrir efectos estacionales más amplios.

3.7.2 Prueba para el año 2017

Con la finalidad de mostrar cuál es el efecto que genera el nuevo modelo propuesto, se harán las pruebas necesarias para verificar su confiablidad haciendo una prueba con los datos y verificando los resultados con los datos reales.

Se analizan los datos proporcionados por la serie hasta diciembre de 2016, y se toma el modelo ARIMA que mejor se adapta a los datos[12].

El mejor modelo es un ARIMA (2,1,1) , este modelo cumple con todos los supuestos para pronosticar certeramente. Sin embargo, el pronóstico tiende a ser volátil como se puede observar en el gráfico siguiente:

Gráfico 3.13. *Pronóstico del desempleo en EMSyS para 2017.*

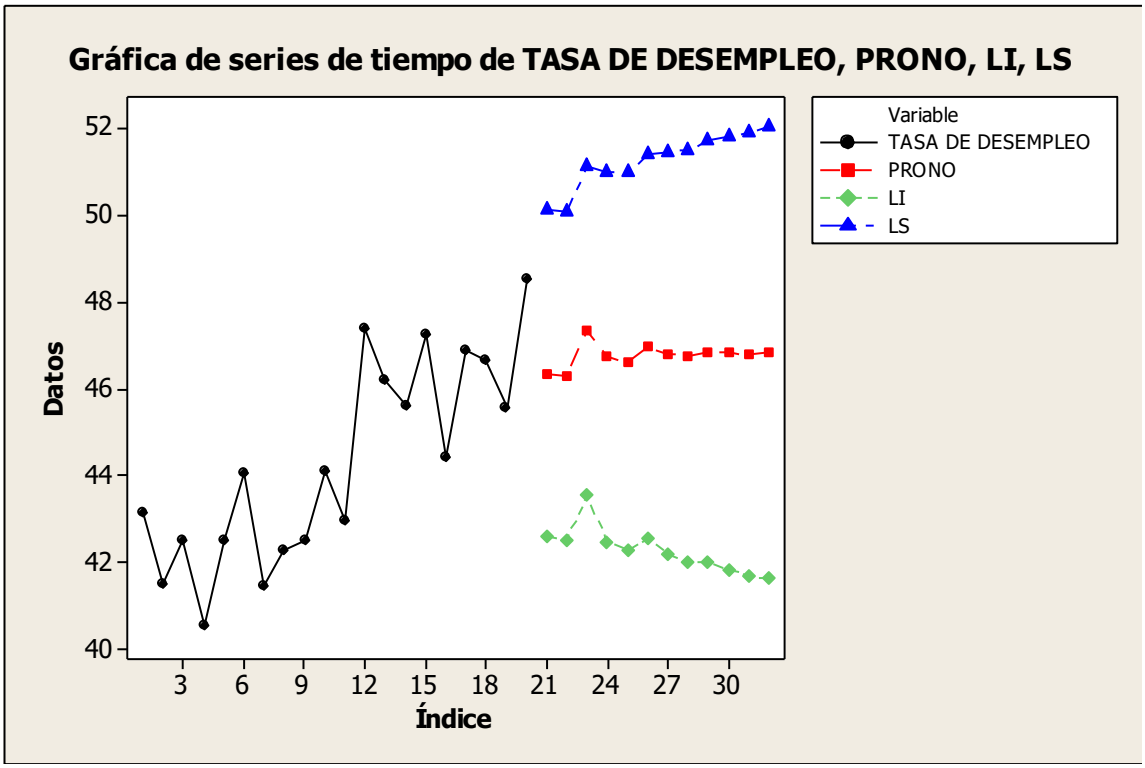


Gráfico que representa la serie de tiempo con datos empíricos y el pronóstico con correspondiente límite superior e inferior.

Fuente: Elaboración Propia

[12] Por la facilidad de los cálculos se toma la función auto.arima del software R

Es decir, sí pronostica certeramente pero el problema se vuelve un poco trivial debido a la abertura del intervalo de confianza. Es por ello que se propone tomar las variables artificiales como un factor temporal además de no aleatorio.

Teniendo las variables artificiales además de la serie original es necesario realizar una regresión. Esta, incluirá los términos AR y MA determinados por el modelo ARIMA además de las variables dummy. En este caso, es necesario tomar dos componentes autoregresivos, uno de promedios móviles y las doce variables artificiales correspondientes a cada mes. Por lo tanto se busca obtener un modelo con la siguiente estructura:

$$y_t = \beta_0 + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varphi_1 \varepsilon_{t-1} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} \quad (3.2)$$

Donde:

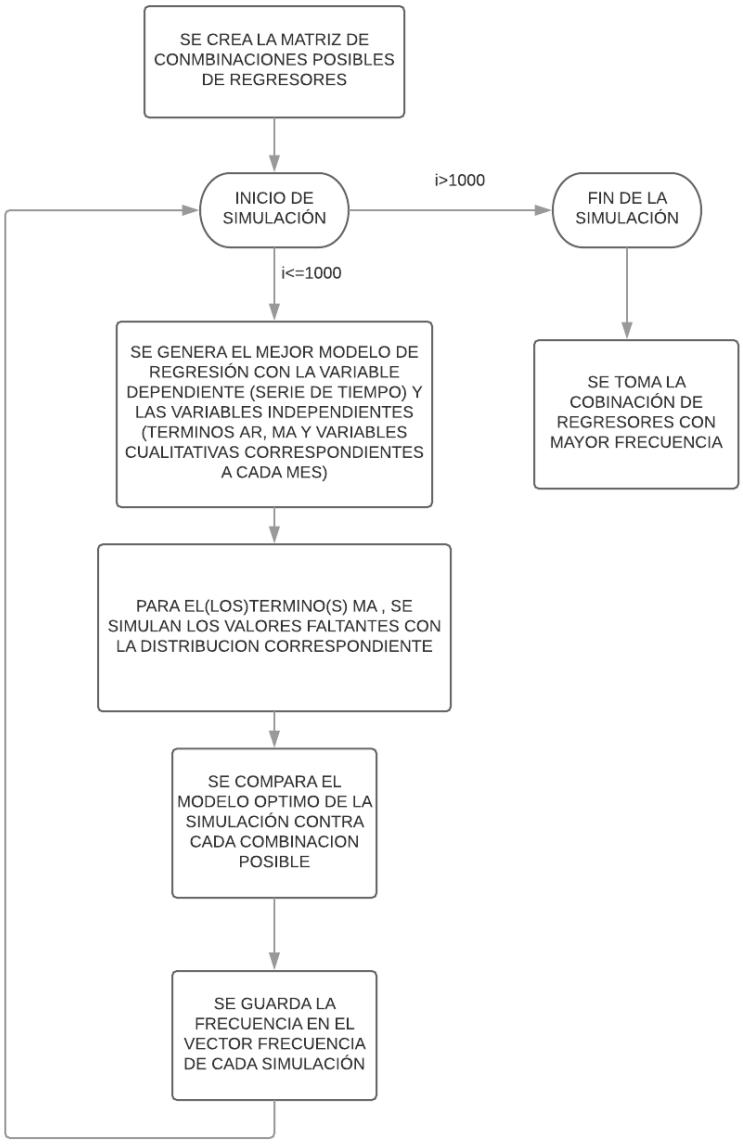
Cada una de las variables x es una variable binaria asociada a cada mes, tomando valor 1 si el dato pertenece al mes correspondiente, es importante hacer notar que se toma como base el mes de enero, es decir, cuando todas las x tienen valor cero se asocia a este mes, esto se hace debido a que debe cuidarse que no exista la multicolinealidad en las variables independientes. Es importante también debido a que dará información clave en la interpretación de los coeficientes de regresión.

Sin embargo, si se desea obtener el valor del pronóstico de este modelo mediante regresión para cada mes del 2017 es necesario tener los datos de la variable independiente para todo el 2017 además de tener los doce valores de promedios móviles. El cálculo del valor de la variable dependiente se toma del pronóstico del modelo ARIMA óptimo. La determinación de los valores faltantes de promedios móviles en sí no es difícil, ya que gracias al supuesto de normalidad de residuales se conocen los parámetros como se distribuyen. En este caso tienen normalidad con media 0.3865 y desviación estándar de 1.867. No obstante, generar estos valores no es un proceso que se deba hacer solamente una vez, es decir, se tiene que generar una simulación y seleccionar el modelo que se repita la mayor cantidad de veces, eso dará la certeza de que se toma el mejor modelo de regresión para explicar la tasa de desempleo. El modelo de regresión debe

estar obligado a tomar en cuenta lo ya determinado por el modelo ARIMA, es decir, dos términos autoregresivos y uno de promedios móviles. Este proceso es mejor si se hace con ayuda de software computacional, en este caso se usará R. Se realizan 1000 simulaciones a través de la estructura mostrada en el gráfico 3.12. Lo primero es utilizar la información brindada por Hyndman & Athanasopoulos (2018) con la paquería llamada “forecast” es posible utilizar una automatización en la selección del modelo ARIMA con la confianza en el cumplimiento de todas los supuestos necesarios para tener el mejor modelo correspondiente a los datos. Este paso es necesario ya que es físicamente imposible verificar cada uno de los supuestos para cada modelo en las 1000 simulaciones.

El proceso de simulación también requerirá hacer pruebas de ajuste asociadas a los pronósticos y residuos, este procedimiento tiende a ser tedioso debido a la gran variedad de distribuciónes de probabilidad, es por ello que se utiliza la paquetería desarrollada por (Belgorodski & Greiner, 2017) en la cual se facilita la comparación de muchas distribucines continuas lo cual ahorra desarrollo de código de programación.

Gráfico 3.12. *Estructura de la simulación.*



Fuente: Elaboración Propia

Una vez realizadas las mil simulaciones se determina que el mejor modelo tomará en cuenta los siguientes parámetros:

$$y_t = \beta_0 + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varphi_1 \varepsilon_{t-1} + \beta_2 x_2 + \beta_5 x_5 + \beta_8 x_8 + \beta_{11} x_{11} \qquad (3.3)$$

Donde se puede hacer notar que la combinación de regresores que más se repite son los que toman en cuenta los parámetros ya determinados por el modelo ARIMA además de

las variables correspondientes a los meses de Marzo, Junio, Septiembre y Diciembre. En este paso se ha determinado cuáles son los parámetros significativos para el modelo. Es importante hacer notar que al realizar las simulaciones no siempre se tendrán los mismos pronósticos ni los mismos coeficientes de los parámetros, es por ello que es necesario tomar los vectores generados por la simulación y evaluarlos. Sin embargo, mediante el código de programación se puede determinar qué distribución tienen estos coeficientes de los regresores como se muestra a continuación:

Cuadro 3.1. *Distribución de los parámetros.*

Coeficiente	Distribución	Media	Desviación Estándar
β_0	Normal	-1.7600046	0.4471071
θ_1	Lognormal	-2.0885149	0.1272101
θ_2	Normal	0.91004958	0.02417234
φ_1	Ninguna	-0.67917857	0.04229843
β_2	Normal	2.4515529	0.01105483
β_5	Normal	3.0180665	0.1034274
β_8	Normal	1.165550	0.109221
β_{11}	Normal	1.6298561	0.1059832

Fuente: Elaboración Propia

Con la información que se muestra en la tabla 3.1 se puede determinar que el modelo en términos generales para el año 2017 cuenta con la siguiente forma:

$$y_t = -1.76 + 0.12 + 0.91y_{t-2} - 0.67\varepsilon_{t-1} + 2.45x_2 + 3.01x_5 + 1.16x_8 + 1.62x_{11} \quad (3.4)$$

En general, todo modelo conlleva un error por lo que lo más adecuado sería representarlo como:

$$y_t = -1.76 + 0.12 + 0.91y_{t-2} - 0.67\varepsilon_{t-1} + 2.45x_2 + 3.01x_5 + 1.16x_8 + 1.62x_{11} + \varepsilon_t \quad (3.5)$$

Antes de analizar a fondo el pronóstico, es necesario contraponer los datos reales del 2017 por lo cual se realiza la prueba. Para ello, es necesario restar al vector pronóstico los datos reales de cada mes. Con esto se generará el vector de residuales el cual necesario analizar. Resultando en lo siguiente:

Cuadro 3.2. *Distribución de los residuales*

Vector Residual del Mes	Distribución	Media	Desviación Estándar
Enero	Normal	-0.5869223	0.1332942
Febrero	Normal	0.3390645	0.1396011
Marzo	Normal	1.190420	1.208265
Abril	Normal	-1.317924	1.202971
Mayo	Normal	-1.284867	1.202971
Junio	Normal	-0.03794239	1.15786715
Julio	Normal	-0.588969	1.253854
Agosto	Normal	-0.6871766	1.1893404
Septiembre	Normal	-0.7508921	1.2096427
Octubre	Normal	-0.5895997	1.3182118

Noviembre	Normal	-0.9855968	1.2511049
Diciembre	Normal	-0.4932296	1.2772140

Fuente: Elaboración Propia

Este análisis para el año 2017 muestra un hecho valorable del modelo, los residuales tienen distribución normal con lo cual cubre el supuesto necesario para confirmar que el modelo de regresión es bueno.

Una vez analizados los residuales, se puede estudiar el comportamiento del pronóstico sobre cada mes, es decir, se tienen mil pronósticos sobre cada mes lo cual sirve como una base fundamental de estudio. Utilizando la prueba de Anderson-Darling y Kolmogorov-Smirnov se demuestra que los vectores del pronóstico tienen distribución normal:

Cuadro 3.3. *Distribución del pronóstico para 2017*

Vector Pronóstico del Mes	Distribución	Media	Desviación Estándar
Enero	Normal	46.0395195	0.1332942
Febrero	Normal	46.9620940	0.1396011
Marzo	Normal	48.819118	1.208265
Abril	Normal	46.001335	1.202971
Mayo	Normal	46.001335	1.202971
Junio	Normal	47.609895	1.157867

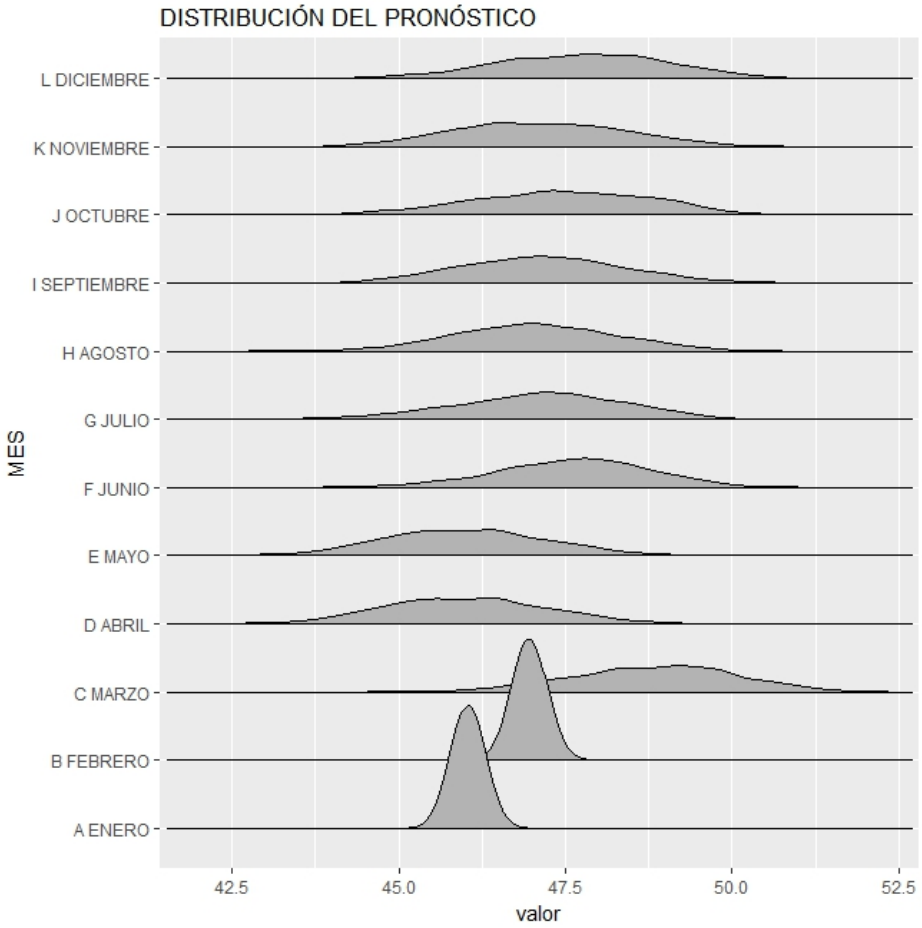
Julio	Normal	47.099793	1.253854
Agosto	Normal	47.04891	1.18934
Septiembre	Normal	47.160801	1.209643
Octubre	Normal	47.427710	1.318212
Noviembre	Normal	47.121425	1.251105
Diciembre	Normal	47.740587	1.277214

Fuente: Elaboración Propia

Al saber que el pronóstico de cada mes generado tiene una distribución normal es un paso muy importante para conocer más acerca del modelo.

Es decir, el pronóstico del modelo original ARIMA tenía la un pronóstico muy volátil como se muestra en el gráfico 3.13. Simplemente una barra de intervalo de confianza en la cual no se podía saber con exactitud la distribución del error sobre cada pronóstico. Mientras que el pronóstico del modelo ARIMA modificado:

Gráfico 3.14. *Pronóstico del desempleo en EMSyS para 2017 con el modelo ARIMA modificado.*



Fuente: Elaboración Propia

Si bien el pronóstico generado por el modelo ARIMA modificado sigue teniendo anchura grande, en este se conoce la distribución que tiene cada pronóstico además de la probabilidad asociada a cada pronóstico además de incluir implícitamente factores temporales cualitativos. Es importante remarcar cómo los pronósticos para Enero y Febrero tienen más precisión sobre la distribución normal correspondiente, esto es debido a que el pronóstico se basa en datos reales como ya se había explicado previamente.

De esta forma, contrastando los datos reales contra los del modelo se tienen las pruebas necesarias para decir que el modelo que se utilizará para el pronóstico de la tasa de desempleo para el 2018 es confiable y acertado.

3.8 Pronóstico para la tasa de desempleo de personas con EMSyS para 2018 con el modelo ARIMA modificado e Implicaciones Económicas

Se determinó en la sección 3.6, el mejor modelo para pronosticar la tasa de desempleo para 2018 es un ARIMA (2,1,2). Sin embargo, como ya se había mencionado, es necesario utilizar variables cualitativas para disminuir la varianza del pronóstico además de utilizar simulación para conocer mejor cómo se distribuyen los valores pronosticados sobre cada mes.

Utilizando la estructura del gráfico 3.12. Se realiza la simulación en el que se determina el modelo de regresión óptima para predecir.

Al calcular el modelo con 1000 simulaciones, se puede hacer notar que la combinación óptima de regresores no toma en cuenta exactamente los dos términos MA del correspondiente modelo ARIMA, es decir, solo toma en cuenta los dos términos autoregresivos y uno de promedios móviles. Es importante remarcar que de las 1000 simulaciones, en 985 toma en cuenta sólo estos y no el segundo término MA. Debido a la contundencia de este dato se ha decidido no obligar al modelo a tomar en cuenta esta última variable. Entonces se tiene la siguiente combinación óptima de regresores:

$$y_t = \beta_0 + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varphi_1 \varepsilon_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_6 x_6 + \beta_{12} x_{12} \qquad (3.6)$$

Una vez conociendo cuáles es la combinación óptima de los mismos, se utiliza la media de los vectores correspondientes a los parámetros para tener una idea aproximada del modelo general:

$$y_t = 1.23 + .39 y_{t-1} + .55 y_{t-2} + .37 \varepsilon_1 + 1.24 x_2 + 2.35 x_3 + 2.48 x_6 + 1.77 x_{12} \qquad (3.7)$$

Este modelo con las medias a simple vista se podría definir simplemente como un modelo matemático que explica la tasa de desempleo para todos los meses del año 2018. Sin embargo, tiene una gran significancia en el impacto de cada uno de ellos. Más adelante se ahondará en ello.

Una vez teniendo el modelo medio, con el cual se puede expresar cómo se desarrollará el pronóstico, es necesario observar los datos que arroja el modelo. Utilizando pruebas se puede determinar que el pronóstico para cada mes tiene una distribución normal como se muestra a continuación:

Cuadro 3.4. *Distribución del pronóstico para 2018*

Vector Pronóstico del Mes	Distribución	Media	Desviación Estándar
Enero	Normal	45.4742	0.0597
Febrero	Normal	45.4977	0.0309
Marzo	Normal	47.9161	0.0693
Abril	Normal	46.7313	0.6592
Mayo	Normal	46.1173	0.7013
Junio	Normal	49.7399	0.6548
Julio	Normal	46.8921	0.6968
Agosto	Normal	46.9290	0.7010
Septiembre	Normal	47.3028	0.6760
Octubre	Normal	47.2140	0.6979

Noviembre	Normal	47.3228	0.6659
Diciembre	Normal	49.2698	0.6438

Este modelo tiene una alta confianza y ha mejoradp el pronóstico del ARIMA tradicional, ya que se puede conocer la distribución sobre cada mes. Sin embargo, es necesario probar que los residuales tienen distribución normal además de varianza constante. No obstante, este paso es muy complicado debido a la pesadez del requerimiento computacional. Se sabe además, que no se tienen los valores reales para el año 2018 es por ello que se toma como base el pronóstico del ARIMA tradicional para verificar que se tenga normalidad de los residuales para este año.

Cuadro 3.5. *Distribución de los residuales para el año 2018*

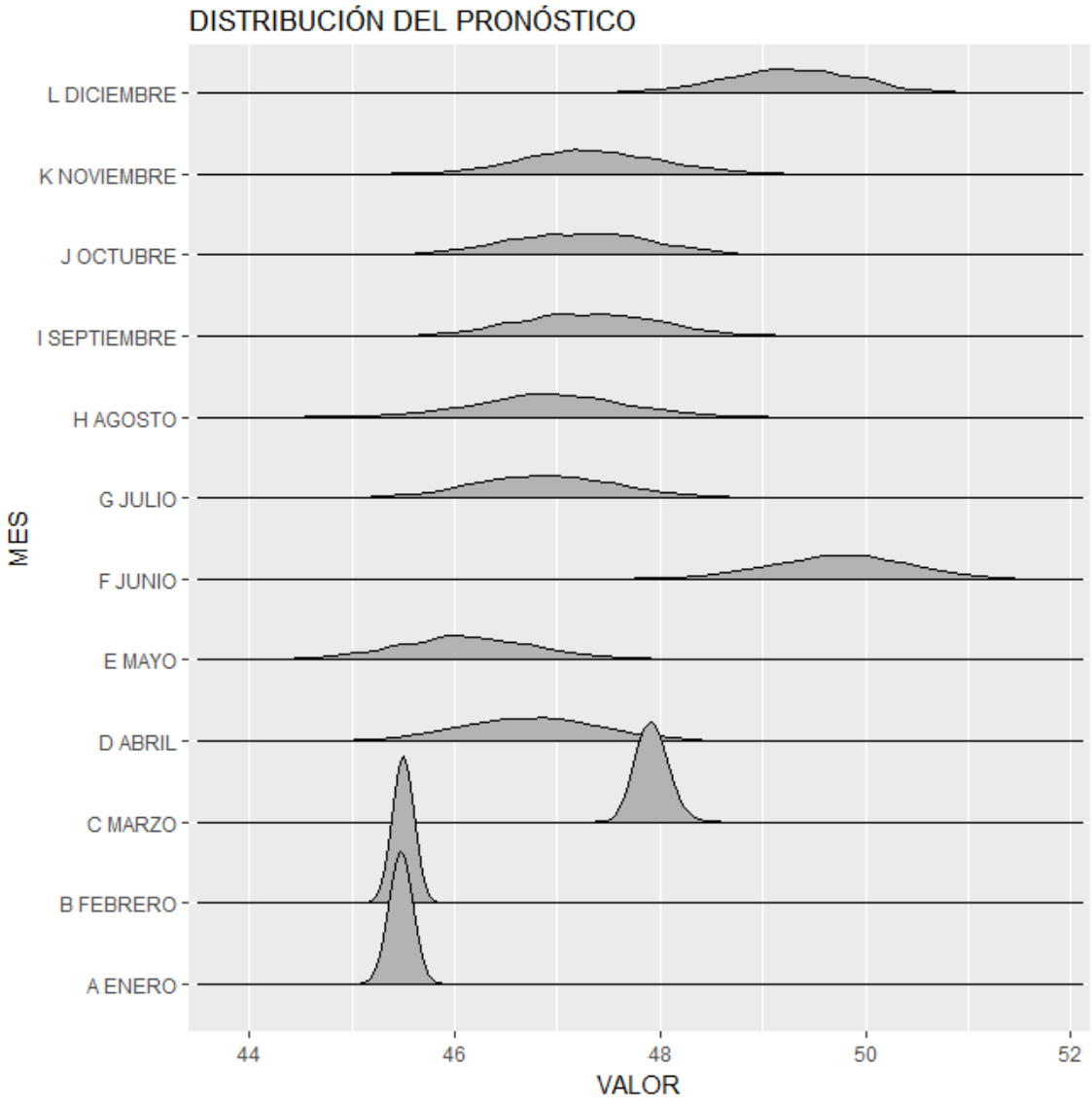
Vector Residual del Mes	Distribución	Media	Desviación Estándar
Enero	Normal	-0.9953	.0597
Febrero	Normal	-1.0039	0.0309
Marzo	Normal	-2.6158	0.0693
Abril	Normal	-1.2784	0.6592
Mayo	Normal	2.1678	0.7013
Junio	Normal	1.2510	0.6587
Julio	Normal	0.0196	0.6968

Agosto	Normal	-3.7923	0.7010
Septiembre	Normal	0.7092	0.6760
Octubre	Normal	0.3180	0.6979
Noviembre	Normal	-2.4572	0.6659
Diciembre	Normal	6.9215	0.6438

A medida que el pronóstico se va a alejando del punto de inicio, es decir, de los datos reales, la tendencia a la volatilidad es mayor. Sin embargo, se mantiene la normalidad sobre los residuales por lo que se puede concluir que el modelo es bueno para pronosticar.

Como se observa el gráfico 3.13. El pronóstico generado con el modelo ARIMA modificado tiene también varianza muy grande. Sin embargo, esta ha sido delimitada por una distribución de probabilidad, en este caso normal. Esto es de gran utilidad ya que se puede conocer con exactitud de donde a donde recorrerá el pronóstico para cada mes.

Gráfico 3.15. *Pronóstico del desempleo en EMSyS para 2018 con el modelo ARIMA modificado.*



3.9 Implicaciones Económicas

Si bien, es muy importante el pronóstico en sí, también cuenta con gran utilidad sacarle provecho al modelo algebraico de series de tiempo empleado como herramienta para predecir. Dicho modelo se expresa a continuación:

$$y_t = 1.23 + .39y_{t-1} + .55y_{t-2} + .37\varepsilon_1 + 1.24x_2 + 2.35x_3 + 2.48x_6 + 1.77x_{12} \quad (3.8)$$

En donde se ve la dependencia de la serie en 8 factores llamadas variables independientes. Donde:

- y_i : Factores asociados al i-ésimo retraso
- ε_i : Factores asociado al i-ésimo término de promedios móviles, es decir, errores del pronóstico pasado.
- x_i : Si el dato se encuentra en el i-ésimo mes del año

Es decir, muestra factores asociados a la variable. Se puede aseverar que el desempleo en México para personas con educación media superior y superior tendrá una base de 1.23%, es decir, es un valor del cual no se bajará en el país. Esto determinado por diversas circunstancias no propias de este estudio. Esta declaración se extrae del factor 1.23 del modelo, el cual no importa ningún factor, se tendrá este valor.

Los componentes interesantes de analizar son los determinados por el mes en el que se encuentra el dato. Se ha tomado como base el mes de Enero debido a que si el dato se encuentra en este mes, todos los valores de las x_i son cero. Es por ello que la interpretación será en base a este mes.

Basados en el modelo se determina que se tiene aproximadamente 1.24% ,2.35%, 2.48%, 1.77% más desempleo en Febrero, Marzo, Junio, Diciembre respectivamente que en Enero. Esto nos da un indicador de cómo se comportará la tasa de desempleo, sin necesidad de ver los datos duros.

Esta herramienta matemática con interpretación económica tiene una ayuda invaluable en por ejemplo, políticas públicas. Es decir, el gobierno trata de minimizar la tasa de desempleo con muchas acciones realizadas, sin embargo, es difícil concluir qué tanto ayudan las mismas a minorizarlo. En cambio, teniendo este modelo se sabe que si en Febrero hubo menos de 2.35% de desempleo que en Enero. Sí se han estado haciendo las cosas de manera correcta.

Conclusiones

Se obtienen las siguientes conclusiones sobre esta trabajo:

El modelo ARIMA (2,1,1) propuesto mediante la metodología Box-Jenkins describe de forma adecuada el comportamiento del porcentaje de desocupación en personas con educación media superior y superior de México .

La calidad de la base de datos es aceptable y el modelo sirve para evaluarla respecto a las nuevas observaciones que se vayan agregando. Sin embargo, el modelo óptimo de series de tiempo para este porcentaje es corto debido a que las predicciones por delante de dos meses, es decir, a partir de marzo de 2018 son inestables debido a la volatilidad de los intervalos de confianza, con esto se prueba la hipótesis plateada al principio del trabajo. No obstante, el modelo ARIMA modificado, que incluye variables cualitativas correspondientes a los meses del año además de los parámetros determinados por la metodología, brinda un pronóstico con su correspondiente nivel de confianza pero brinda además la distribución de cada pronóstico, con lo cual se puede conocer de manera más certera el valor que obtendrá. Con este modelo se puede confirmar gráfica y probabilísticamente cómo a medida que el pronóstico se aleja de los valores reales empíricos, el pronóstico tiene una varianza mayor.

Una serie que se define linealmente por el valor de la observación anterior además de depender de un mayor número de parámetros tiene una mayor estabilidad del pronóstico debido a tener como base fundamental la serie original, es por ello, que si el modelo concluyente tiene como dependencia dos parámetros de observaciones anteriores además de errores del modelo en sí, se tiene un pronóstico más inestable.

Sin importar la volatilidad del pronóstico, el modelo es muy confiable por lo cual puede ser utilizado para mostrar la tendencia general de este porcentaje para todo el año 2018. Debido a esta volatilidad del pronóstico, se puede concluir que la hipótesis del trabajo se comprueba.

Además de todo, el modelo es una forma clara de medición de las políticas aplicadas en el país durante este año, con esta herramienta se puede mostrar si una reforma es de real beneficio para las personas con educación media superior y superior o simplemente un gasto innecesario más del gobierno en turno.

Es importante hacer notar que el 2018 es un año que tuvo cambio de presidente por lo cual es gran magnitud tener vigilancia y control de uno de los indicadores claves de la economía mexicana.

En general, el modelo y el pronóstico se tomaron de una base de datos corta a comparación de las existentes en países con mayor cultura de la medición de los indicadores económicos de la economía. La ENOE realizada en México mensualmente es una encuesta todavía joven que si bien es una forma confiable de conocer la ocupación y el empleo en el país todavía no brinda las herramientas practicas necesarias para tener un modelo con menores errores y mayor certeza pero lo obtenido es una buena base para conocer cómo se compone el porcentaje de personas con educación media superior y superior en México.

Teniendo los datos pronosticados de la tasa de desocupación para personas con EMSyS se tiene una base fundamental para observar cómo se desarrolló el bienestar de las familias en la actividad económica durante el 2018. Este estudio muestra la capacidad suficiente para aplicarse en los años venideros.

ANEXOS

Anexo 1 . Código de programación

```
library("forecast")
library(MASS)
fr=1
AA<-c(1)
BB<-c(1)
CC<-c(1)
PP<-c(0)
DD<-c(0,1)
EE<-c(0,1)
FFF<-c(0,1)
GG<-c(0,1)
HH<-c(0,1)
II<-c(0,1)
JJ<-c(0,1)
KK<-c(0,1)
LL<-c(0,1)
MM<-c(0,1)
NN<-c(0,1)
OO<-c(0,1)
EXPAND<-expand.grid(AA,BB,CC,PP,DD,EE,FFF,GG,HH,II,JJ,KK,LL,MM,NN,OO)
library(readr)
datos <- read_csv("C:/Users/JulioTBL/Desktop/10°/MODELO 2018/datos.csv")
seriedetiempo<-c(30.146588,31.60615819,datos$Y)
seriedetiempo
modeloarima<-auto.arima(seriedetiempo)
summary(modeloarima)
pronostico<-forecast(modeloarima,h=14)
pronostico$mean
VECTORFRECUENCIAS<-rep(0,4096)
```

```
aa<-c(seriedetiempo,pronostico$mean)
vector_serie<-
c(datos$Y,pronostico$mean[1],pronostico$mean[2],pronostico$mean[3],pronostico$mean[4],
      pronostico$mean[5],pronostico$mean[6],pronostico$mean[7],pronostico$mean[8],

pronostico$mean[9],pronostico$mean[10],pronostico$mean[11],pronostico$mean[12])
vector_ar1<-c(datos$`YT-
1`,datos$Y[154],pronostico$mean[1],pronostico$mean[2],pronostico$mean[3],pronostico$mean[4],

pronostico$mean[5],pronostico$mean[6],pronostico$mean[7],pronostico$mean[8],
          pronostico$mean[9],pronostico$mean[10],pronostico$mean[11])
vector_ar2<-c(datos$`YT-
2`,datos$Y[153],datos$Y[154],pronostico$mean[1],pronostico$mean[2],pronostico$mean[3],pronostico$mean[4],

pronostico$mean[5],pronostico$mean[6],pronostico$mean[7],pronostico$mean[8],
          pronostico$mean[9],pronostico$mean[10])
vector_enero<-c(datos$ENERO,0,0,0,0,0,0,0,0,0,0,0,0)
vector_febrero<-c(datos$FEBRERO,0,1,0,0,0,0,0,0,0,0,0,0)
vector_marzo<-c(datos$MARZO,0,0,1,0,0,0,0,0,0,0,0,0)
vector_abril<-c(datos$ABRIL,0,0,0,1,0,0,0,0,0,0,0,0)
vector_mayo<-c(datos$MAYO,0,0,0,0,1,0,0,0,0,0,0,0)
vector_junio<-c(datos$JUNIO,0,0,0,0,0,1,0,0,0,0,0,0)
vector_julio<-c(datos$JULIO,0,0,0,0,0,0,1,0,0,0,0,0)
vector_agosto<-c(datos$AGOSTO,0,0,0,0,0,0,0,1,0,0,0,0)
vector_septiembre<-c(datos$SEPTIEMBRE,0,0,0,0,0,0,0,0,1,0,0,0)
vector_octubre<-c(datos$OCTUBRE,0,0,0,0,0,0,0,0,0,1,0,0)
vector_noviembre<-c(datos$NOVIEMBRE,0,0,0,0,0,0,0,0,0,0,1,0)
vector_dic<-c(datos$DICIEMBRE,0,0,0,0,0,0,0,0,0,0,0,1)
```

```
##############################################################################
############
for (i in 1:1000){
  z1<- rnorm(1, 0.3877, 1.908)
  z2<-rnorm(2, 0.3877, 1.908)
  z3<-rnorm(9, 0.3877, 1.908)
  z4<-rnorm(8, 0.3877, 1.908)

  zz1<-c (z1,modeloarima$residuals,z3)
  zz2<-c (z2,modeloarima$residuals,z4)
  fit <-
step(lm(vector_serie~vector_ar1+vector_ar2+zz1+zz2+vector_enero+vector_febrero+ve
ctor_marzo+
                vector_abril+vector_mayo+vector_junio+vector_julio+
                vector_agosto+vector_septiembre+vector_octubre+
                vector_noviembre+vector_dic))

  A=0
  B=0
  C=0
  Dd=0
  D=0
  E=0
  FF=0
  G=0
  H=0
  I=0
  J=0
  K=0
  L=0
  M=0
```

```
N=0
O=0
xc<-colnames(fit$model)
xc
length(xc)
po<-ifelse(xc %in% c("vector_ar1"),1,0)
A=sum(po)
po<-ifelse(xc %in% c("vector_ar2"),1,0)
B=sum(po)
po<-ifelse(xc %in% c("zz1"),1,0)
C=sum(po)
po<-ifelse(xc %in% c("zz2"),1,0)
Dd=sum(po)
po<-ifelse(xc %in% c("vector_enero"),1,0)
D=sum(po)
po<-ifelse(xc %in% c("vector_febrero"),1,0)
E=sum(po)
po<-ifelse(xc %in% c("vector_marzo"),1,0)
FF=sum(po)
po<-ifelse(xc %in% c("vector_abril"),1,0)
G=sum(po)
po<-ifelse(xc %in% c("vector_mayo"),1,0)
H=sum(po)
po<-ifelse(xc %in% c("vector_junio"),1,0)
I=sum(po)
po<-ifelse(xc %in% c("vector_julio"),1,0)
J=sum(po)
po<-ifelse(xc %in% c("vector_agosto"),1,0)
K=sum(po)
po<-ifelse(xc %in% c("vector_septiembre"),1,0)
L=sum(po)
```

```
po<-ifelse(xc %in% c("vector_octubre"),1,0)
M=sum(po)
po<-ifelse(xc %in% c("vector_noviembre"),1,0)
N=sum(po)
po<-ifelse(xc %in% c("vector_dic"),1,0)
O=sum(po)
vector_combinaciones<-c(A,B,C,Dd,D,E,FF,G,H,I,J,K,L,M,N,O)
vector_combinaciones
for(fr in 1:4096){
  mn1<-EXPAND$Var1[fr]
  mn2<-EXPAND$Var2[fr]
  mn3<-EXPAND$Var3[fr]
  mn4<-EXPAND$Var4[fr]
  mn5<-EXPAND$Var5[fr]
  mn6<-EXPAND$Var6[fr]
  mn7<-EXPAND$Var7[fr]
  mn8<-EXPAND$Var8[fr]
  mn9<-EXPAND$Var9[fr]
  mn10<-EXPAND$Var10[fr]
  mn11<-EXPAND$Var11[fr]
  mn12<-EXPAND$Var12[fr]
  mn13<-EXPAND$Var13[fr]
  mn14<-EXPAND$Var14[fr]
  mn15<-EXPAND$Var15[fr]
  mn16<-EXPAND$Var16[fr]
  if(mn1==A&&mn2==B&&mn3==C&&mn4==Dd&&mn5==D&&mn6==E&&
     mn7==FF&&mn8==G&&mn9==H&&mn10==I&&mn11==J&&mn12==K&&
     mn13==L&&mn14==M&&mn15==N&&mn16==O)
  {

    VECTORFRECUENCIAS[fr]=VECTORFRECUENCIAS[fr]+1
```

```
   }
  }

}

VECTORFRECUENCIAS
max(VECTORFRECUENCIAS)
View(VECTORFRECUENCIAS )
sum(VECTORFRECUENCIAS)
write.csv(VECTORFRECUENCIAS, file="datanuevaS.csv") #guardamos en un archivo
CSV.

MODELO_MEJOR<-
c(EXPAND$Var1[2087],EXPAND$Var2[2087],EXPAND$Var3[2087],EXPAND$Var4[208
7],EXPAND$Var5[2087],

EXPAND$Var6[2087],EXPAND$Var7[2087],EXPAND$Var8[2087],EXPAND$Var9[2087]
,EXPAND$Var10[2087],

EXPAND$Var11[2087],EXPAND$Var12[2087],EXPAND$Var13[2087],EXPAND$Var14[
2087],EXPAND$Var15[2087],EXPAND$Var16[2087]
MODELO_MEJOR

BANDE=0
VECPE<-1:2000
VECPFEB<-1:2000
VECPMAR<-1:2000
VECPABR<-1:2000
VECPMAYO<-1:2000
VECPJUN<-1:2000
VECPJUL<-1:2000
```

```
VECPAG<-1:2000
VECPSP<-1:2000
VECPOC<-1:2000
VECPNOV<-1:2000
VECPDIC<-1:2000

tt1<-1:2000
tt2<-1:2000
tt3<-1:2000
tt4<-1:2000
tt5<-1:2000
tt6<-1:2000
tt7<-1:2000
tt8<-1:2000


for (qr in 1:2000)
{


  z1<- rnorm(1, 0.3877, 1.908)
  z2<-rnorm(2, 0.3877, 1.908)
  z3<-rnorm(9, 0.3877, 1.908)
  z4<-rnorm(8, 0.3877, 1.908)

  zz1<-c (z1,modeloarima$residuals,z3)
  zz2<-c (z2,modeloarima$residuals,z4)
```

```
 fit <-
step(lm(vector_serie~vector_ar1+vector_ar2+zz1+zz2+vector_enero+vector_febrero+ve
ctor_marzo+
                vector_abril+vector_mayo+vector_junio+vector_julio+
                vector_agosto+vector_septiembre+vector_octubre+
                vector_noviembre+vector_dic))
 SS<-colnames(fit$model)


 po<-ifelse(SS %in% c("vector_ar1"),1,0)
 A=sum(po)
 po<-ifelse(SS %in% c("vector_ar2"),1,0)
 B=sum(po)
 po<-ifelse(SS %in% c("zz1"),1,0)
 C=sum(po)
 po<-ifelse(SS %in% c("zz2"),1,0)
 DD=sum(po)
 po<-ifelse(SS %in% c("vector_enero"),1,0)
 D=sum(po)
 po<-ifelse(SS %in% c("vector_febrero"),1,0)
 E=sum(po)
 po<-ifelse(SS %in% c("vector_marzo"),1,0)
 FF=sum(po)
 po<-ifelse(SS %in% c("vector_abril"),1,0)
 G=sum(po)
 po<-ifelse(SS %in% c("vector_mayo"),1,0)
 H=sum(po)
 po<-ifelse(SS %in% c("vector_junio"),1,0)
 I=sum(po)
 po<-ifelse(SS %in% c("vector_julio"),1,0)
 J=sum(po)
```

```
  po<-ifelse(SS %in% c("vector_agosto"),1,0)
  K=sum(po)
  po<-ifelse(SS %in% c("vector_septiembre"),1,0)
  L=sum(po)
  po<-ifelse(SS %in% c("vector_octubre"),1,0)
  M=sum(po)
  po<-ifelse(SS %in% c("vector_noviembre"),1,0)
  N=sum(po)
  po<-ifelse(SS %in% c("vector_dic"),1,0)
  O=sum(po)

  if(A== 1&& B==1 && C==1&& DD==0 && D==0&& E== 1&& FF==1&&G== 0&&
H==0&&I== 1&&J== 0&&K== 0&&L== 0&&
     M==0&&N== 0&&O== 1)
  {
    BANDE<-BANDE+1
    VECPE[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[155]+
      fit$coefficients[3]*vector_ar2[155]+fit$coefficients[4]*vector_zz1[155]

    VECPFEB[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[156]+

fit$coefficients[3]*vector_ar2[156]+fit$coefficients[4]*vector_zz1[156]+fit$coefficients[5]*
vector_febrero[156]


    VECPMAR[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[157]+

fit$coefficients[3]*vector_ar2[157]+fit$coefficients[4]*zz1[157]+fit$coefficients[6]*vector_
marzo[157]
```

```
VECPABR[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[158]+
  fit$coefficients[3]*vector_ar2[158]+fit$coefficients[4]*zz1[158]

VECPMAYO[BANDE]<- fit$coefficients[1]+fit$coefficients[2]*vector_ar1[159]+
  fit$coefficients[3]*vector_ar2[159]+fit$coefficients[4]*zz1[159]

VECPJUN[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[160]+
fit$coefficients[3]*vector_ar2[160]+fit$coefficients[4]*zz1[160]++fit$coefficients[7]*vector
_junio[160]

VECPJUL[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[161]+
  fit$coefficients[3]*vector_ar2[161]+fit$coefficients[4]*zz1[161]

VECPAG[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[162]+
  fit$coefficients[3]*vector_ar2[162]+fit$coefficients[4]*zz1[162]

VECPSP[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[163]+
  fit$coefficients[3]*vector_ar2[163]+fit$coefficients[4]*zz1[163]

VECPOC[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[164]+
  fit$coefficients[3]*vector_ar2[164]+fit$coefficients[4]*zz1[164]

VECPNOV[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[165]+
```

```
      fit$coefficients[3]*vector_ar2[165]+fit$coefficients[4]*zz1[165]


    VECPDIC[BANDE]<-fit$coefficients[1]+fit$coefficients[2]*vector_ar1[166]+

fit$coefficients[3]*vector_ar2[166]+fit$coefficients[4]*zz1[166]+fit$coefficients[8]*vector_
dic[166]


    tt1[BANDE]<-fit$coefficients[1]
    tt2[BANDE]<-fit$coefficients[2]
    tt3[BANDE]<-fit$coefficients[3]
    tt4[BANDE]<-fit$coefficients[4]
    tt5[BANDE]<-fit$coefficients[5]
    tt6[BANDE]<-fit$coefficients[6]
    tt7[BANDE]<-fit$coefficients[7]
    tt8[BANDE]<-fit$coefficients[8]



  }
}
P1<-1:BANDE
P2<-1:BANDE
P3<-1:BANDE
P4<-1:BANDE
P5<-1:BANDE
P6<-1:BANDE
P7<-1:BANDE
P8<-1:BANDE
P9<-1:BANDE
```

```
P10<-1:BANDE
P11<-1:BANDE
P12<-1:BANDE


TT1<-1:BANDE
TT2<-1:BANDE
TT3<-1:BANDE
TT4<-1:BANDE
TT5<-1:BANDE
TT6<-1:BANDE
TT7<-1:BANDE
TT8<-1:BANDE
for (sas in 1:BANDE)
{
  P1[sas]=VECPE[sas]
  P2[sas]=VECPFEB[sas]
  P3[sas]=VECPMAR[sas]
  P4[sas]=VECPABR[sas]
  P5[sas]=VECPMAYO[sas]
  P6[sas]=VECPJUN[sas]
  P7[sas]=VECPJUL[sas]
  P8[sas]=VECPAG[sas]
  P9[sas]=VECPSP[sas]
  P10[sas]=VECPOC[sas]
  P11[sas]=VECPNOV[sas]
  P12[sas]=VECPDIC[sas]
  TT1[sas]=tt1[sas]
  TT2[sas]=tt2[sas]
  TT3[sas]=tt3[sas]
  TT4[sas]=tt4[sas]
```

```
  TT5[sas]=tt5[sas]
  TT6[sas]=tt6[sas]
  TT7[sas]=tt7[sas]
  TT8[sas]=tt8[sas]

}


library("rriskDistributions")
mean(P1)
write.csv(P1, file="PRONOSTICCO1.csv") #guardamos en un archivo CSV.
PP1<-P1[-c(1)]
PP2<-P2[-c(1)]
PP3<-P3[-c(1)]
PP4<-P4[-c(1)]
PP5<-P5[-c(1)]
PP6<-P6[-c(1)]
PP7<-P7[-c(1)]
PP8<-P8[-c(1)]
PP9<-P9[-c(1)]
PP10<-P10[-c(1)]
PP11<-P11[-c(1)]
PP12<-P12[-c(1)]

TTT1<-TT1[-c(1)]
TTT2<-TT2[-c(1)]
TTT3<-TT3[-c(1)]
TTT4<-TT4[-c(1)]
TTT5<-TT5[-c(1)]
TTT6<-TT6[-c(1)]
TTT7<-TT7[-c(1)]
```

```
TTT8<-TT8[-c(1)]
mean(TT8)
fit.cont(PP1)
fit.cont(PP2)
fit.cont(PP3)
fit.cont(PP4)
fit.cont(PP5)
fit.cont(PP6)
fit.cont(PP7)
fit.cont(PP8)
fit.cont(PP9)
fit.cont(PP10)
fit.cont(PP11)
fit.cont(PP12)
fit.cont(TTT1)
fit.cont(TTT2)
fit.cont(TTT3)
fit.cont(TTT4)
fit.cont(TTT5)
fit.cont(TTT6)
fit.cont(TTT7)
fit.cont(TTT8)

RESIDUALENERO<-fit.cont(PP1-vector_serie[143])
RESIDUALFEB<-fit.cont(PP2-vector_serie[144])
RESIDUALMARZ<-fit.cont(PP3-vector_serie[145])
RESIDUALABR<-fit.cont(PP4-vector_serie[146])
RESIDUALMAYO<-fit.cont(PP5-vector_serie[147])
RESIDUALJUN<-fit.cont(PP6-vector_serie[148])
RESIDUALJUL<-fit.cont(PP7-vector_serie[149])
RESIDUALAGOS<-fit.cont(PP8-vector_serie[150])
```

```
RESIDUALSEP<-fit.cont(PP9-vector_serie[151])
RESIDUALOCT<-fit.cont(PP10-vector_serie[152])
RESIDUALNOV<-fit.cont(PP11-vector_serie[153])
RESIDUALDIC<-fit.cont(PP12-vector_serie[154])
PRONOSTICO2018<-
data.frame(PP1,PP2,PP3,PP4,PP5,PP6,PP7,PP8,PP9,PP10,PP11,PP12)
install.packages("ggjoy")
library(ggjoy)
library(readr)
PRONOSTICO_2018 <- read_csv("C:/Users/JulioTBL/Desktop/10°/MODELO
2018/pronostico2018.csv")
X11();
ggplot(PRONOSTICO_2018,aes(x=VALOR,y=MES))+geom_joy2() +
  scale_x_continuous(expand = c(0.01, 0))+
  labs(title = 'DISTRIBUCIÓN DEL PRONÓSTICO')
```

Anexo 2 . Pruebas de Normalidad

H0: La muestra proviene de una distribución normal.

H1: La muestra no proviene de una distribución normal.

Para pruebas de normalidad siempre se plantean así las hipótesis.

El nivel de significancia que se trabajará es de 0.05. Alfa=0.05

Criterio de Decisión

Si P < Alfa Se rechaza Ho

Si p >= Alfa No se rechaza Ho

Pruebas sobre pronóstico 2018

```
                logL      AIC      BIC Chisq(value) Chisq(p) AD(value)    H(A
D) KS(value)        H(KS)
Normal        2745.84 -5487.69 -5476.52        38.85     0.22      0.76 reject
ed      0.02 not rejected
Cauchy        2354.74 -4705.48 -4694.31       571.35     0.00     30.77 reject
ed      0.08     rejected
Logistic      2720.12 -5436.24 -5425.08        76.10     0.00      3.29 reject
ed      0.03     rejected
Exponential  -9460.87 18923.75 18929.33   1141758.77     0.00    898.53 reject
ed      0.63     rejected
Chi-square   -6230.34 12462.68 12468.27    218819.79     0.00    749.10     NU
LL      0.51     rejected
Uniform          NULL     NULL     NULL          Inf     0.00       Inf     NU
LL      0.06     rejected
Gamma         2745.82 -5487.63 -5476.47        38.87     0.22      0.76 reject
ed      0.02 not rejected
Weibull       2601.06 -5198.12 -5186.96       207.79     0.00     19.11 reject
ed      0.06     rejected
F           -12592.28 25188.55 25199.72   5631280.20     0.00   1101.41     NU
LL      0.70     rejected
Student     -14073.96 28149.92  28155.5  11976444.18     0.00   1949.89     NU
LL      0.84     rejected

Chosen continuous distribution is: Normal (norm)
Fitted parameters are:
       mean          sd
45.47423533  0.05978347

                logL      AIC      BIC Chisq(value) Chisq(p) AD(value)    H(A
D) KS(value)    H(KS)
Normal        4040.49 -8076.97 -8065.81        62.84        0      2.74 reject
ed      0.04 not rejected
Cauchy        3652.56 -7301.12 -7289.95       560.54        0     33.04 reject
ed      0.09 rejected
Logistic      4017.72 -8031.45 -8020.28        85.56        0      4.58 reject
ed      0.04 rejected
Exponential  -9461.89 18925.78 18931.36   2228811.26        0    899.65 reject
ed      0.63 rejected
Chi-square   -6230.82 12463.65 12469.23    428536.93        0    754.45     NU
LL      0.51 rejected
Uniform          NULL     NULL     NULL          Inf        0       Inf     NU
LL      0.06 rejected
```

```
Gamma          4040.35 -8076.69 -8065.53        63.01        0      2.76 reject
ed      0.04 rejected
Weibull        3998.25 -7992.49 -7981.33       142.27        0      9.57 reject
ed      0.04 rejected
F             -12593.5    25191 25202.17  10986688.63        0   1101.00     NU
LL      0.70 rejected
Student      -14075.19 28152.39 28157.97  23364185.90        0   1950.01     NU
LL      0.84 rejected

Chosen continuous distribution is: Normal (norm)
Fitted parameters are:
       mean          sd
45.49773855  0.03092431
```

Bibliografía

Box, G. E. (1970). *Time series analysis, forecasting and control*

Barcelata Chávez, H. (s.f.). Crítica a la teoría y política neoliberal del empleo. *Universidad Veracruzana*, 1.

Belgorodski, N., & Greiner, M. (24 de Marzo de 2017). Fitting Distributions to Given Data or Known Quantiles.

Burgos Flores, B., & López Montes, K. (2010). La situación del mercado laboral de profesionistas. *Revista de la educación superior*.

Cran. (11 de Abril de 2018). Obtenido de https://cran.r-project.org/web/packages/forecast/forecast.pdf

Diario Oficial de la Federación. (20 de mayo de 2013). Obtenido de http://www.dof.gob.mx/nota_detalle_popup.php?codigo=5326569

Gujarati, D., & Down C, P. (2010). *Econometría* (Quinta ed.). Estados Unidos: McGraw Hill. Recuperado el 13 de Mayo de 2017

Harvey, D. (2005). *Breve historia del Neoliberalismo.* Oxford University Press Canada.

Heath, J. (2012). La Importancia de la Tasa de Desempleo. *Pulso Económico*.

Hyndman, R., & Athanasopoulos, G. (11 de Abril de 2018). Forecasting Functions for Time Series and Linear Models.

INEGI. (14 de Noviembre de 2017). Obtenido de http://www.inegi.org.mx/saladeprensa/boletines/2017/enoe_ie/enoe_ie2017_11.pdf

Jiménez Jiménez, J. (2005). Las políticas del empleo en México y el desarrollo regional. *Aportes*.

O'Conell, R., Bowerman, B., & Koehler, A. (2005). *Forecasting, Time Series, and Regression: An Applied Approach.* South-Western College Pub .

Organización Internacional del Trabajo. (2016). *Perspectivas sociales y del empleo en el mundo.* Ginebra.

(2017). *Panorama General de los Indicadores Laborales en México.* Universidad de Guanajuato, Guanajuato.

Romero, M. (2008). *Gobierno de Puerto Rico.* Obtenido de Departamento del Trabajo y Recursos Humanos: http://www.trabajo.pr.gov/pdf/Estadisticas/GT/Grupo%20Trabajador/2010/El%20DESEMPLEO.pdf

Rosales, R. (s.f.). *Universidad De Los Andes.* Obtenido de https://economia.uniandes.edu.co/files/profesores/ramon_rosales_alvarez/docs/econometria2/Salidas%20y%20Ejercicios/EJC202220Metodologa20Box20-20Jenkins.pdf

Ruíz Nápoles, P., & Ordaz Díaz, J. (s.f.). Evolución reciente del empleo y el desempleo en México. *ECONOMÍA*, 91-92.

Secretaría del Trabajo y Previsión Social. (Mayo de 2008). *INEGI.* Obtenido de http://www.inegi.org.mx/rne/docs/Pdfs/Mesa1/20/OscarOrtiz.pdf

Situación del Empleo en México. (2013). *Contaduría Pública*, 44-46.

Trejo Reyes , S. (1972). El desempleo en México; Características Generales. *Comercio Exterior*, 730-738.